

Verso un'intelligenza autonoma

La proposta teorica di Yann LeCun tra architetture cognitive, fondamenti filosofici e implicazioni etiche

Prefazione metodologica

Il presente saggio nasce da una domanda che attraversa tutto il dibattito contemporaneo sull'intelligenza artificiale: è possibile costruire macchine che comprendano il mondo, o siamo condannati a sistemi che lo simulano senza abitarlo? Yann LeCun, tra i più influenti ricercatori del campo, ha posto questo interrogativo al centro della propria riflessione con una radicalità che pochi altri si sono concessi. La sua risposta — articolata in un corpus di paper, interventi pubblici, dispute accademiche e posizioni filosofiche — è al tempo stesso una critica e un programma: critica al paradigma dominante dei modelli linguistici autoregressivi, programma per un'architettura che aspiri a qualcosa di qualitativamente diverso da ciò che il semplice scaling di quei modelli può produrre.

Il rischio, nella trattazione di proposte così ambiziose, è duplice. Da un lato si profila la deferenza agiografica che trasforma il ricercatore in visionario e le sue ipotesi in certezze acquisite. Dall'altro incombe la critica semplificatoria che riduce un programma scientifico complesso a wishful thinking o a strategia di posizionamento istituzionale. Questa monografia intende evitare entrambe le trappole adottando un metodo che si potrebbe definire di esposizione critica simpatetica: si prende sul serio la proposta di LeCun nei suoi dettagli tecnici e nelle sue implicazioni più profonde, si ricostruisce il territorio concettuale che la circonda con la precisione che il soggetto impone, e si esercita il giudizio critico con la medesima cura con cui si è esposta la tesi.

Una nota su un aspetto che permea l'intera trattazione. LeCun ha spesso descritto i modelli linguistici di grandi dimensioni come sistemi che "non capiscono nulla", applicando — implicitamente e talvolta esplicitamente — la metafora del pappagallo stocastico coniata da Emily Bender e colleghi nel 2021. È doveroso affrontare questa caratterizzazione con la stessa serietà con cui si affronta il resto della sua proposta: essa solleva questioni filosoficamente legittime, ma incorpora anche semplificazioni che rischiano di oscurare fenomeni genuinamente complessi. La monografia dedica a questo tema una sezione specifica, nel tentativo di portare chiarezza su un dibattito che si svolge troppo spesso ai confini tra argomentazione scientifica e posizionamento retorico.

La struttura segue una progressione dal particolare al generale e poi di nuovo al particolare: si parte dal contesto tecnico, si sale al piano delle implicazioni cognitive e filosofiche, e si ridiscende alla valutazione critica delle prospettive concrete. Sul piano formale, le derivazioni matematiche vengono esposte in notazione plaintext con simboli Unicode, integrate nel corpo del testo in modo da rendere solidale l'apparato tecnico alla riflessione discorsiva, anziché relegarlo in un supplemento separato. I riferimenti bibliografici sono raccolti in coda, con rimandi interni per le precisazioni più specifiche.

1. Introduzione: il momento critico dell'intelligenza artificiale

Pochi momenti nella storia della scienza sono stati tanto carichi di aspettative quanto quello che stiamo attraversando nel campo dell'intelligenza artificiale. Il dispiegamento su scala globale di sistemi come GPT-4, Claude, Gemini e i loro successori ha portato nella sfera pubblica la percezione — ora euforica, ora angosciata — che qualcosa di fondamentale stia mutando nelle capacità cognitive delle macchine. Eppure, al di sotto di questa convergenza percettiva, il disaccordo tra i ricercatori non è mai stato così profondo, e le fratture investono questioni di principio che toccano la natura stessa della comprensione, la struttura della conoscenza, il significato di parole come "capire", "ragionare", "intendere".

LeCun sostiene che i sistemi attualmente dominanti — i Large Language Models, d'ora in poi LLM — siano strutturalmente incapaci di raggiungere qualcosa che assomigli all'intelligenza generale, non per difetto di scala o di dati, ma per ragioni architetturali profonde che nessun incremento quantitativo può sanare. Il problema, nella sua diagnosi, non risiede nel numero di parametri né nella dimensione dei dataset: è il paradigma stesso a essere inadeguato. I modelli autoregressivi apprendono a predire il prossimo simbolo in una sequenza ottimizzando una funzione che non richiede né favorisce la costruzione di modelli causali del mondo. Possono diventare straordinariamente abili in questo compito — e i risultati empirici ne sono testimonianza eloquente — ma questa abilità, nella visione lecuniana, non si traduce in comprensione autentica e non può farlo per ragioni di principio.

La proposta che LeCun contrappone al paradigma dominante si articola intorno a tre concetti che si sostengono vicendevolmente: l'apprendimento autosupervisionato in spazio latente, il Modello di Mondo come struttura centrale dell'agente intelligente, e JEPa come architettura concreta che implementa questi principi. L'ambizione è costruire sistemi che apprendano principalmente dall'osservazione non supervisionata del mondo fisico — come fanno gli animali e i bambini — sviluppando rappresentazioni interne capaci di supportare pianificazione, ragionamento causale e adattamento robusto a situazioni inedite.

La posta in gioco, come si vedrà, trascende qualunque preferenza architettonica. È in discussione la concezione di cosa significhi "capire", di come si formi la conoscenza del mondo, di quale sia il rapporto tra linguaggio e pensiero, di quali rischi e responsabilità comporti lo sviluppo di sistemi potenzialmente autonomi. Per questo la monografia si muove su piani distinti ma connessi, convinta che la separazione tra il tecnico e il filosofico sia in questo caso non solo artificiosa ma epistemicamente dannosa: chi si chiude nel tecnico manca le implicazioni più profonde della proposta; chi si limita al filosofico rischia di discutere di astrazioni senza fondamento nell'architettura reale dei sistemi in questione.

2. Profilo intellettuale di Yann LeCun

2.1 Formazione e traiettoria scientifica

Yann LeCun nasce nel 1960 a Soisy-sous-Montmorency, nei pressi di Parigi. La sua formazione avviene dapprima all'ESIEE Paris, poi all'Université Pierre-et-Marie-Curie, dove consegue il dottorato nel 1987 sotto la supervisione di Maurice Milgram. È in questo periodo che iniziano a prendere forma le intuizioni che condurranno al suo primo grande contributo: l'applicazione sistematica dell'algoritmo di retropropagazione del gradiente alle reti neurali convoluzionali per il riconoscimento di strutture visive.

L'esperienza ai Bell Labs tra il 1988 e il 1996 è formativa in duplice senso: scientificamente, perché è lì che sviluppa LeNet — la rete convoluzionale che elaborava milioni di assegni bancari al giorno anticipando di vent'anni la rivoluzione del deep learning — e intellettualmente, perché l'ambiente dei Bell Labs, tra i più fertili della ricerca informatica del Novecento, gli offre l'esposizione a una varietà di discipline che rimarrà un tratto costante del suo approccio. Nel 2003 passa alla New York University, dove co-fonda il Center for Data Science del Courant Institute. Nel 2013 viene nominato Chief AI Scientist di Meta, dove ha guidato FAIR (Fundamental AI Research), il laboratorio di ricerca fondamentale del gruppo. Nel 2025 ha lasciato anche Meta per fondare una propria startup con l'obiettivo di intraprendere l'ambiziosa e pluriennale strada verso il nuovo paradigma AI.

2.2 Le reti convoluzionali e il riconoscimento visivo: fondamenti formali

La contribuzione di LeCun alle reti neurali convoluzionali non è un dettaglio implementativo: costituisce un salto concettuale la cui portata si comprende appieno soltanto attraverso la sua formalizzazione matematica. LeNet introduce — o più esattamente formalizza per la prima volta in modo operativo e sperimentalmente validato — il principio dell'invarianza spaziale nella percezione visiva artificiale.

L'operazione fondamentale è la convoluzione discreta bidimensionale. Data una feature map di input $f : \mathbb{Z}^2 \rightarrow \mathbb{R}$ e un filtro $g : \mathbb{Z}^2 \rightarrow \mathbb{R}$ a supporto limitato, la feature map di output è:

$$(f \star g)(m, n) = \sum_k \sum_l f(k, l) \cdot g(m - k, n - l)$$

In un'implementazione a passo finito su tensori discreti, con kernel di dimensione $F \times F$, stride s e zero-padding p , il valore dell'uscita nella posizione (i, j) per il canale di output c è:

$$h_c[i, j] = \sum_{c'} \sum_{u=0}^{F-1} \sum_{v=0}^{F-1} x_{c'}[s \cdot i + u - p, s \cdot j + v - p] \cdot w_c[c', u, v] + b_c$$

dove x è il tensore di input con canali c' , w_c è il tensore dei pesi del filtro per il canale di output c , e b_c è il bias scalare. La dimensione spaziale della feature map di output è data da:

$$O = \lfloor (I - F + 2p) / s \rfloor + 1$$

con I dimensione spaziale dell'input. Il kernel w_c è condiviso su tutte le posizioni spaziali (weight sharing): questa è l'essenza del principio di invarianza traslazionale imposto dalla struttura, che riduce drasticamente il numero di parametri rispetto a uno strato completamente connesso e costringe il modello a rilevare lo stesso pattern indipendentemente dalla sua localizzazione nell'immagine.

Il legame con la neurobiologia non è ornamentale. I lavori di David Hubel e Torsten Wiesel sulla corteccia visiva primaria (V1) dei gatti, premiati con il Nobel nel 1981, avevano documentato che neuroni di tipo simple cell rispondono selettivamente a orientamenti preferiti di bordi in posizioni specifiche del campo visivo, mentre neuroni di tipo complex cell mostrano risposte analoghe ma invarianti rispetto a piccole traslazioni dello stimolo. La convoluzione discreta è, in prima approssimazione, un modello computazionale delle simple cell; l'operazione di max-pooling che tipicamente segue ogni strato convoluzionale — formalmente:

$$h_pool[i, j] = \max_{(u,v) \in \mathcal{N}(i,j)} h[u, v]$$

dove $\mathcal{N}(i,j)$ denota un intorno locale di raggio r centrato in (i,j) — è il corrispondente computazionale delle complex cell, realizzando l'invarianza locale a piccoli spostamenti del pattern preferito.

In una rete profonda con L strati convoluzionali, la rappresentazione del dato x al livello l è:

$$h^{(l)} = \sigma(W^{(l)} \star h^{(l-1)} + b^{(l)})$$

dove σ è una nonlinearity elemento per elemento — tipicamente $\text{ReLU}(z) = \max(0, z)$ nelle formulazioni post-2012 — e $h^{(0)} = x$ è l'input. La composizione iterata di L trasformazioni di questo tipo produce una

funzione parametrica $f_{\theta} : \mathbb{R}^{(H \times W \times C)} \rightarrow \mathbb{R}^d$ che mappa un'immagine di dimensioni spaziali $H \times W$ con C canali in un vettore di rappresentazione d -dimensionale. La proprietà cruciale di questa gerarchia è che i livelli bassi catturano strutture locali ad alta frequenza (bordi, gradienti di colore, angoli) mentre i livelli alti catturano strutture semantiche globali (parti di oggetti, oggetti interi, scene complesse). Tale organizzazione non è imposta manualmente dall'ingegnere: emerge dall'ottimizzazione della perdita di addestramento attraverso retropropagazione del gradiente — il che significa che la rete apprende autonomamente a strutturare le proprie rappresentazioni in livelli di astrazione crescente.

L'algoritmo di backpropagation, nella sua applicazione alle reti convoluzionali, merita una trattazione formale rigorosa poiché costituisce il motore computazionale dell'intero apprendimento profondo moderno. Data una funzione di perdita scalare $\mathcal{L}(\theta)$, l'obiettivo è calcolare $\partial \mathcal{L} / \partial W^{(l)}$ per ogni strato $l = L, L-1, \dots, 1$ attraverso l'applicazione iterata della regola di derivazione composta. Definendo il segnale di errore retropropagato al livello l come:

$$\delta^{(l)} = \partial \mathcal{L} / \partial z^{(l)}$$

dove $z^{(l)} = W^{(l)} \star h^{(l-1)} + b^{(l)}$ è la pre-attivazione, la ricorrenza di backpropagation è:

$$\delta^{(l)} = (W^{(l+1)})^{\text{rot}180} \star \delta^{(l+1)} \odot \sigma'(z^{(l)})$$

in cui $\odot$ indica il prodotto di Hadamard (elemento per elemento), $\sigma'(z^{(l)})$ è la derivata element-wise della nonlinearity valutata nel punto $z^{(l)}$, e $(W^{(l+1)})^{\text{rot}180}$ è il kernel ruotato di 180° che trasforma la convoluzione in una correlazione crociata trasposta (operazione equivalente alla convoluzione piena nell'equazione del gradiente). Il gradiente rispetto ai pesi è poi:

$$\partial \mathcal{L} / \partial W^{(l)} = \delta^{(l)} \star h^{(l-1)}$$

dove $\star$ indica questa volta la correlazione crociata tra il segnale d'errore e le attivazioni del livello precedente. Questo risultato, che LeCun e collaboratori avevano derivato nel loro paper fondamentale del 1989 e poi esposto sistematicamente nel contributo del 1998 su *Proceedings of the IEEE*, stabilisce le basi algoritmiche su cui si regge l'intero paradigma del deep learning.

L'intuizione gerarchica che ne emerge — le buone rappresentazioni come struttura a livelli di astrazione crescente appresa per ottimizzazione — è il filo conduttore che dall'architettura LeNet porta, attraverso trasformazioni concettuali profonde, alla proposta JEPa. In entrambi i casi, il principio organizzante è che la qualità di una rappresentazione si misura non dalla fedeltà con cui riproduce l'input grezzo, ma dalla capacità di catturare le strutture causalmente rilevanti ignorando le variazioni irrilevanti.

— un principio che nell'era delle reti convoluzionali si chiamava invarianza traslazionale e nell'era di JEPA si chiama predizione in spazio latente.

2.3 Il posizionamento nel dibattito contemporaneo sull'AGI

LeCun è tra i pochi ricercatori di primo piano ad aver assunto posizioni pubbliche esplicite e sistematiche sul futuro dell'intelligenza artificiale, distinguendosi tanto dagli "accelerazionisti" — convinti che il semplice scaling delle architetture attuali condurrà all'AGI entro tempi brevi — quanto dagli "allarmisti" — persuasi che la comparsa di superintelligenze incontrollabili sia imminente e costituisca una minaccia esistenziale per l'umanità. La sua posizione è, in un certo senso, doppiamente scomoda: nega sia l'ottimismo ingenuo del primo campo sia la drammatizzazione del secondo, e viene accusata da entrambi i versanti di distorcere la realtà per ragioni di posizionamento strategico o di difesa del proprio programma di ricerca.

La sua tesi può essere articolata in tre affermazioni che si implicano vicendevolmente. La prima: i modelli linguistici attuali, per quanto impressionanti nelle loro prestazioni su benchmark linguistici, non sono sulla traiettoria dell'intelligenza generale, perché l'ottimizzazione della verosimiglianza su sequenze di token non converge verso la comprensione causale del mondo a prescindere dalla scala raggiunta. La seconda: l'intelligenza genuina — intesa come capacità di pianificazione robusta, ragionamento causale e adattamento flessibile a contesti fisici nuovi — richiede la costruzione di modelli del mondo fisico che i sistemi puramente linguistici non possono sviluppare attraverso il solo addestramento su testo. La terza: i rischi reali dell'intelligenza artificiale non derivano da un'eccessiva intelligenza dei sistemi ma dalla loro inadeguatezza — da agenti che ottimizzano surrogati sbagliati degli obiettivi umani senza comprendere il contesto in cui operano, producendo danni per incompetenza più che per malizia.

Questo posizionamento lo ha messo in aperto conflitto con vecchi collaboratori come Geoffrey Hinton e Yoshua Bengio — con i quali condivide il Premio Turing del 2018, il riconoscimento più prestigioso dell'informatica teorica — che nel corso degli ultimi anni hanno espresso allarme crescente circa i rischi esistenziali dello sviluppo dell'AI. Il dibattito tra queste figure, che si svolge sia in paper accademici sia in forum pubblici e interviste, non è riducibile a una disputa tecnica su architetture o funzioni di perdita: riflette differenze profonde nella valutazione epistemica del rischio in condizioni di incertezza radicale, nel rapporto tra ricerca di base e responsabilità sociale, e nella concezione di cosa significhi agire con prudenza di fronte a sviluppi tecnologici il cui impatto futuro non può essere stimato con certezza.

3. La critica al paradigma autoregressivo

3.1 Architettura formale degli LLM: il Transformer

I modelli linguistici di grandi dimensioni nella loro forma attuale sono sistemi per la stima della distribuzione di probabilità condizionale su sequenze di simboli discreti. L'architettura dominante — il Transformer, introdotto da Vaswani e collaboratori in "Attention Is All You Need" (2017) — ha imposto la propria struttura come standard de facto per una gamma straordinariamente vasta di compiti. La comprensione della critica di LeCun richiede una familiarità non superficiale con questa architettura, e la sua trattazione formale merita dunque uno spazio adeguato.

Sia $\mathcal{V}$ un vocabolario finito di cardinalità $|\mathcal{V}|$ (tipicamente dell'ordine di 50.000 token per modelli come GPT-4), e sia $\mathbf{x} = (x_1, x_2, \dots, x_T) \in \mathcal{V}^T$ una sequenza di lunghezza T . Attraverso una matrice di embedding $E \in \mathbb{R}^{(|\mathcal{V}| \times d_{\text{model}})}$, ciascun token x_t viene proiettato in un vettore continuo $\mathbf{e}_t = E[x_t] \in \mathbb{R}^{d_{\text{model}}}$. A questi vettori si sommano encoding posizionali $\mathbf{p}_t \in \mathbb{R}^{d_{\text{model}}}$ — nella formulazione originale di Vaswani et al., sinusoidali fissi:

$$\mathbf{p}_t[2k] = \sin(t / 10000^{(2k/d_{\text{model}})}) \quad \mathbf{p}_t[2k+1] = \cos(t / 10000^{(2k/d_{\text{model}})})$$

per $k = 0, 1, \dots, d_{\text{model}}/2 - 1$. Questi encoding permettono al modello di distinguere posizioni diverse nella sequenza senza ricorrere alla ricorrenza o alla convoluzione. In architetture più recenti (RoPE, ALiBi) gli encoding posizionali vengono integrati direttamente nel meccanismo di attenzione in forme che migliorano la generalizzazione a sequenze più lunghe di quelle viste in addestramento.

La sequenza di vettori arricchiti con informazione posizionale $\mathbf{X}^{(0)} \in \mathbb{R}^{(T \times d_{\text{model}})}$, con $\mathbf{X}^{(0)}_t = \mathbf{e}_t + \mathbf{p}_t$, attraversa poi N blocchi identici. Ogni blocco si compone di due sottomoduli: Multi-Head Self-Attention (MHSA) e una rete feed-forward posizione-per-posizione (FFN), ciascuno seguito da normalizzazione di strato e connessione residuale.

Il meccanismo di Multi-Head Self-Attention con H teste parallele opera come segue. Per ciascuna testa $h \in \{1, \dots, H\}$, si definiscono le proiezioni:

$$\mathbf{Q}_h = \mathbf{X} \mathbf{W}_h^Q \in \mathbb{R}^{(T \times d_k)} \quad \mathbf{K}_h = \mathbf{X} \mathbf{W}_h^K \in \mathbb{R}^{(T \times d_k)} \quad \mathbf{V}_h = \mathbf{X} \mathbf{W}_h^V \in \mathbb{R}^{(T \times d_v)}$$

dove $\mathbf{W}_h^Q, \mathbf{W}_h^K \in \mathbb{R}^{(d_{\text{model}} \times d_k)}$ e $\mathbf{W}_h^V \in \mathbb{R}^{(d_{\text{model}} \times d_v)}$ sono matrici di proiezione apprendibili, con $d_k = d_v = d_{\text{model}} / H$. Il prodotto scalare normalizzato tra query e chiavi produce una matrice di punteggi di attenzione grezzi:

$$\mathbf{S}_h = \mathbf{Q}_h \mathbf{K}_h^T / \sqrt{d_k} \in \mathbb{R}^{(T \times T)}$$

Il fattore di scala $1/\sqrt{d_k}$ è introdotto per contrastare il fenomeno della saturazione della softmax: quando d_k è grande, il prodotto scalare $\mathbf{q} \cdot \mathbf{k}$ tende ad assumere valori assoluti elevati spingendo la softmax in regioni a gradiente quasi nullo, rallentando la convergenza dell'addestramento. La softmax viene poi applicata riga per riga su $\mathbf{S}_h$, con una maschera

causale M che pone a $-\infty$ le posizioni future (per garantire che la predizione del token t dipenda soltanto da $x_1, \dots, x_{t-1}$):

$$A_h = \text{softmax}(S_h + M) \in \mathbb{R}^{(T \times T)}$$

dove $M_{\{ij\}} = 0$ se $j \leq i$, $-\infty$ se $j > i$. Le righe di A_h sommano a 1 per costruzione, e $A_h[i,j]$ rappresenta il peso con cui il token in posizione j contribuisce alla rappresentazione contestualizzata del token in posizione i . L'uscita della testa h è:

$$\text{head}_h = A_h V_h \in \mathbb{R}^{(T \times d_v)}$$

Le H teste vengono concatenate e riproiettate:

$$\text{MHSA}(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_H) W^O \in \mathbb{R}^{(T \times d_{\text{model}})}$$

con $W^O \in \mathbb{R}^{(H \cdot d_v \times d_{\text{model}})}$. La complessità computazionale di MHSA è $O(T^2 \cdot d_{\text{model}})$ per la moltiplicazione matriciale $Q_h K_h^T$, il che costituisce un collo di bottiglia per sequenze lunghe e ha motivato una proliferazione di meccanismi di attenzione a complessità ridotta.

Ogni blocco include anche una rete feed-forward applicata in modo identico a ogni posizione:

$$\text{FFN}(x_t) = W_2 \max(0, W_1 x_t + b_1) + b_2$$

con $W_1 \in \mathbb{R}^{(d_{\text{ff}} \times d_{\text{model}})}$, $W_2 \in \mathbb{R}^{(d_{\text{model}} \times d_{\text{ff}})}$, e $d_{\text{ff}} = 4 d_{\text{model}}$ tipicamente. I due sottomoduli sono organizzati con connessioni residuali e normalizzazione:

$$\tilde{X}^{(l)} = \text{LayerNorm}(X^{(l-1)} + \text{MHSA}(X^{(l-1)})) \quad X^{(l)} = \text{LayerNorm}(\tilde{X}^{(l)} + \text{FFN}(\tilde{X}^{(l)}))$$

La normalizzazione di strato per un vettore $x \in \mathbb{R}^d$ è:

$$\text{LayerNorm}(x) = \gamma \odot (x - \mu) / \sqrt{(\sigma^2 + \epsilon)} + \beta$$

con $\mu = (1/d) \sum_j x_j$, $\sigma^2 = (1/d) \sum_j (x_j - \mu)^2$, $\gamma, \beta \in \mathbb{R}^d$ parametri apprendibili di scala e bias, e ϵ un termine di stabilità numerica dell'ordine di 10^{-5} . Le connessioni residuali permettono al gradiente di fluire inalterato attraverso molti strati, risolvendo il problema del gradiente che svanisce che affliggeva le architetture pre-residuali.

La funzione di perdita di addestramento è la cross-entropy mediata sull'intero corpus $\mathcal{D}$:

$$\mathcal{L}(\theta) = - (1/|\mathcal{D}|) \sum_{\{x \in \mathcal{D}\}} (1/T_x) \sum_{t=1}^{T_x} \log P_{\theta}(x_t \mid x_1, \dots, x_{t-1})$$

Questa perdita è l'entropia crociata tra la distribuzione empirica dei dati e la distribuzione del modello, ed è equivalente alla divergenza di Kullback-Leibler tra le due distribuzioni a meno di una costante che non dipende dai parametri:

$$\mathcal{L}(\theta) = H(P_{\text{data}}, P_{\theta}) = H(P_{\text{data}}) + \text{KL}(P_{\text{data}} \parallel P_{\theta})$$

ove $H(P_{\text{data}}) = -\mathbb{E}_{\{x \sim P_{\text{data}}\}}[\log P_{\text{data}}(x)]$ è l'entropia della distribuzione vera dei dati (costante rispetto a θ). Minimizzare $\mathcal{L}(\theta)$ equivale pertanto a minimizzare $\text{KL}(P_{\text{data}} \parallel P_{\theta})$, ovvero a fare convergere la distribuzione del modello verso quella empirica dei dati.

3.2 Le leggi di scala e la scaling hypothesis

Le scaling laws documentate da Kaplan e collaboratori (2020) e riformulate con maggiore precisione da Hoffmann e collaboratori nel lavoro su Chinchilla (2022) hanno mostrato che la perdita di cross-entropy L di un modello Transformer segue, in un ampio regime di scale, una legge di potenza nelle variabili N (numero di parametri) e D (numero di token di addestramento):

$$L(N, D) \approx (N_c / N)^{\alpha_N} + (D_c / D)^{\alpha_D} + L_{\text{irr}}$$

dove N_c , D_c , α_N , α_D sono costanti empiriche che dipendono dall'architettura e dalla natura dei dati, e L_{irr} è la perdita irriducibile — corrispondente all'entropia del processo generativo dei dati stessi, il limite inferiore che nessun modello può varcare indipendentemente dalla propria dimensione. Kaplan et al. stimarono $\alpha_N \approx 0.076$ e $\alpha_D \approx 0.095$ per modelli Transformer su testo in lingua inglese; Hoffmann et al. mostrarono che i modelli precedenti (GPT-3, Gopher) erano sistematicamente sotto-addestrati rispetto alla loro dimensione, derivando che per un budget di calcolo fisso C (misurato in operazioni floating point) il rapporto ottimale è:

$$N_{\text{opt}} \propto C^{0.5}, D_{\text{opt}} \propto C^{0.5}$$

ovvero il numero ottimale di token di addestramento è circa uguale, in ordine di grandezza, al numero di parametri, e non venti volte inferiore come si era empiricamente osservato nelle versioni precedenti. Questa analisi ha guidato il dimensionamento di modelli come Llama e i successivi, con conseguenze pratiche molto rilevanti sull'efficienza dell'addestramento.

LeCun non contesta la validità empirica di queste leggi nel regime in cui sono state osservate. La sua obiezione è più radicale: anche se le leggi di scala si mantenessero valide per $N \rightarrow \infty$ e $D \rightarrow \infty$, l'ottimizzazione di $L(N, D)$ non convergerebbe verso l'intelligenza generale, perché la funzione obiettivo è un surrogato inadeguato delle capacità cognitive che ci interessano. La perdita L misura quanto bene il modello predice la distribuzione dei token nel corpus di addestramento; non misura la capacità

di ragionamento causale robusto, di pianificazione in ambienti fisici nuovi, di generalizzazione sistematica a situazioni che richiedono la composizione di concetti in modi non visti durante l'addestramento. Ridurre L non implica migliorare queste capacità, perché esse non sono codificate nella funzione obiettivo.

La struttura logica di questo argomento è la seguente. Sia $\mathbb{C}_L$ l'insieme delle capacità acquisite da un modello che ottimizza $L(N,D)$ con parametri arbitrariamente grandi e dati arbitrariamente abbondanti. Sia $\mathbb{C}_{AGI}$ l'insieme delle capacità necessarie per l'intelligenza generale (secondo una qualunque definizione operativa ragionevole). L'affermazione di LeCun è che $\mathbb{C}_L \subsetneq \mathbb{C}_{AGI}$ — ovvero che esiste almeno una capacità richiesta dall'intelligenza generale che non può essere acquisita attraverso l'ottimizzazione della cross-entropy su token di testo, indipendentemente dalla scala. Questa è un'affermazione empiricamente falsificabile, e la sua verifica richiede benchmark progettati per misurare le capacità in questione — il che conduce direttamente al problema della valutazione che si discuterà nel capitolo 12.

3.3 Meccanismo di self-attention come memoria associativa moderna

Per comprendere la critica lecuniana a un livello più profondo, è illuminante rileggere il meccanismo di attenzione Transformer dal punto di vista della teoria delle reti di Hopfield moderne, stabilita da Ramsauer e collaboratori (2020). Questi autori hanno mostrato che un singolo strato di self-attention con softmax è matematicamente equivalente a una rete di Hopfield moderna con funzione energia:

$$E(\xi, X) = -\text{LSE}(\beta, X \xi) + \frac{1}{2} \xi^T \xi + (1/\beta) \log N + \frac{1}{2} M$$

dove $\text{LSE}(\beta, z) = (1/\beta) \log \sum_i \exp(\beta z_i)$ è il log-sum-exp, $\xi \in \mathbb{R}^d$ è il pattern query da recuperare, $X \in \mathbb{R}^{(N \times d)}$ è la matrice delle chiavi (pattern memorizzati), e M è il numero di pattern. La regola di aggiornamento di questa rete — che minimizza E iterativamente — è:

$$\xi^{(t+1)} = X^T \text{softmax}(\beta X \xi^{(t)})$$

Questa espressione coincide esattamente con l'operazione di attenzione softmax: il token query ξ viene aggiornato come combinazione lineare pesata delle chiavi in X , con pesi dati dalla softmax del prodotto scalare. La capacità di memoria di questa rete — il numero massimo di pattern recuperabili senza errore — scala esponenzialmente con la dimensione d del vettore di stato, a differenza della rete di Hopfield classica (Hopfield, 1982) in cui la capacità era proporzionale a d .

Questa interpretazione ha implicazioni dirette per la critica di LeCun. I Transformer sono, da questa prospettiva, sistemi di recupero associativo estremamente potenti: apprendono a recuperare e ricombinare, date query provenienti dal contesto corrente,

i pattern più pertinenti tra quelli incontrati durante l'addestramento. Questa operazione è di straordinaria efficacia per compiti che richiedono il recupero e la combinazione di conoscenze esplicite, ma strutturalmente fragile per compiti che richiedono la costruzione di rappresentazioni genuinamente nuove attraverso ragionamento causale — ragionamento che deve produrre conclusioni non interpolabili dai pattern del corpus di addestramento. LeCun argomenta che la pianificazione in ambienti fisici inediti, la comprensione di situazioni con causalità nuove, e la generalizzazione robusta a distribuzioni out-of-training appartengono a questa seconda categoria.

3.4 Il problema della comprensione versus la fluenza generativa

Il nucleo filosofico della critica lecuniana è la distinzione tra comprensione e fluenza generativa — una distinzione che richiama il celebre esperimento mentale della Stanza Cinese proposto da John Searle nel 1980 e, più in generale, il problema dell'intenzionalità in filosofia della mente, ma che LeCun traduce in termini operativi concreti, identificando classi di compiti in cui i modelli linguistici mostrano pattern di fallimento sistematici incompatibili con una comprensione genuina del mondo fisico.

Sul piano formale, il problema può essere articolato come segue. Un modello linguistico $P_0(x_t \mid x_{\leq t})$ apprende a stimare la distribuzione condizionale dei token nel corpus, che riflette le regolarità statistiche del testo prodotto da agenti umani incorporati nel mondo. Queste regolarità includono, in forma compressa e mediata, informazioni sulla struttura causale del mondo fisico: la distribuzione dei token che seguono "la sfera è stata lanciata verso il basso" differisce da quella dei token che seguono "la sfera è stata appoggiata su un piano", e questa differenza riflette indirettamente la fisica della gravità. Un modello che impara queste distribuzioni acquisisce una certa codifica indiretta delle regolarità fisiche.

Il punto critico è tuttavia che questa codifica è dipendente dal contesto linguistico in cui è stata acquisita e dalla struttura statistica del corpus. Siano $f_L : \mathcal{V}^* \rightarrow \mathbb{R}^{(d_L)}$ la funzione di rappresentazione latente di un LLM e $f_P : \mathcal{X}_{\text{phys}} \rightarrow \mathbb{R}^{(d_P)}$ la funzione di rappresentazione di un sistema che osserva e modella direttamente il mondo fisico attraverso dati sensoriali. La struttura geometrica dello spazio latente di f_L riflette le correlazioni statistiche del testo e le co-occorrenze lessicali; quella di f_P dovrebbe riflettere le invarianze causali del mondo fisico — traiettorie di oggetti, conservazione della quantità di moto, permanenza degli oggetti nell'occlusione. LeCun argomenta che per compiti che richiedono ragionamento causale robusto su scenari fisici nuovi, la struttura di f_P è epistemicamente superiore a quella di f_L , e che questa superiorità è strutturale, non eliminabile con maggiori quantità di dati testuali.

La distinzione operativa può essere precisata attraverso il concetto di generalizzazione sistematica, introdotto da Fodor e Pylyshyn (1988) e testato empiricamente nei benchmark SCAN (Lake e Baroni, 2018), COGS (Kim e Linzen, 2020) e COMPOSITIONAL GENERALIZATION più recenti. Un sistema mostra generalizzazione sistematica se è capace di comprendere e produrre costruzioni linguistiche o pianificazioni fisiche che combinano concetti noti in modi mai visti

durante l'addestramento — ovvero se le sue rappresentazioni sono genuine composizioni di costituenti, non tabelle di consultazione di insiemi di input. I Transformer di grandi dimensioni mostrano generalizzazione sistematica significativamente limitata rispetto agli esseri umani in questi benchmark, sebbene con variazioni che dipendono dalla dimensione del modello, dalla modalità di prompting e dalla presenza di esempi in-context.

Un banco di prova particolarmente rivelatore è costituito dai compiti di ragionamento fisico intuitivo: situazioni in cui un agente deve prevedere l'evoluzione di una scena fisica — cosa accade se si rimuove un supporto, dove cadrà un oggetto lanciato obliquamente, come si sposterà un liquido in un contenitore inclinato. Su questi compiti, i modelli linguistici mostrano prestazioni che degradano significativamente rispetto alle aspettative generate dalla loro fluenza verbale: possono descrivere correttamente le leggi della fisica di base ma fallire nel ragionarne in modo robusto su istanze concrete che variano anche minimamente dai casi esemplificati nei dati di addestramento. Questo pattern — prestazioni solide nelle varianti coperte dal training, degradazione nelle varianti che richiedono composizione o trasferimento — è, per LeCun, la firma di un sistema che ha appreso correlazioni statistiche piuttosto che strutture causali.

3.5 La metafora del pappagallo stocastico: genesi, struttura e limiti

La locuzione "stochastic parrot" compare per la prima volta in "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" di Emily M. Bender, Timnit Gebru, Angelina McMillan-Major e Shmargaret Shmitchell (2021), presentato alla conferenza FAccT 2021. Il paper, divenuto uno dei documenti più citati e contestati nella storia recente della ricerca sull'AI, non nasce come tesi sulla filosofia della mente ma come critica socio-tecnica: il suo fuoco principale è sui rischi ambientali (il costo computazionale e energetico dell'addestramento di modelli di grandissima scala), sociali (l'amplificazione e la normalizzazione di bias presenti nei corpus di addestramento), ed epistemici (la tendenza del campo a scambiare la fluenza linguistica per comprensione autentica).

La metafora del pappagallo stocastico cattura l'idea che un LLM sia fondamentalmente un sistema che ricombina in modi statisticamente plausibili — campionati da una distribuzione di probabilità — pattern linguistici assorbiti durante l'addestramento, senza alcuna comprensione del significato delle espressioni prodotte. L'"stocastico" specifica che questa ricombinazione è campionata da una distribuzione, non deterministica; il "pappagallo" evoca un imitatore di suoni privo di semantica.

È importante notare che LeCun non è l'autore di questa metafora, sebbene l'abbia adottata e amplificata nelle sue critiche agli LLM. Questa distinzione è rilevante perché le motivazioni di Bender et al. e quelle di LeCun sono parzialmente in disaccordo: le prime sottolineano i rischi sociali e ambientali, il secondo insiste sull'inadeguatezza cognitiva e la distanza strutturale dall'intelligenza genuina. Il fatto che la stessa immagine retorica venga mobilitata per motivazioni diverse è già un segnale della sua ambiguità concettuale.

3.6 Una difesa epistemica dei modelli linguistici di grandi dimensioni

La metafora del pappagallo stocastico è retoricamente efficace ma epistemicamente insufficiente, e merita un esame critico approfondito non come apologia aprioristica degli LLM, ma come esercizio di precisione concettuale indispensabile per qualunque valutazione fondata.

Il primo problema è che la metafora presuppone la risposta alla domanda a cui pretende di rispondere. Affermare che un sistema "produce suoni senza capire" implica già sapere cosa significhi "capire" — che è esattamente la questione in disputa. Se si definisce la comprensione come la capacità di produrre output linguisticamente coerenti, contestualmente appropriati e causalmente sensati in una varietà di domini, allora non è affatto chiaro che i migliori LLM attuali non comprendano nel senso della definizione. Se invece si definisce la comprensione come la presenza di qualia esperienziali, di consapevolezza soggettiva, di intenzionalità in senso brentiano, allora la metafora è corretta — ma si applica per lo stesso argomento a qualunque sistema fisico descrivibile in termini computazionali, il che la rende inutile come criterio discriminante.

Il secondo problema riguarda le capacità emergenti. I modelli linguistici di grandi dimensioni mostrano sistematicamente abilità non previste dai loro progettisti e non presenti nelle versioni di scala minore: ragionamento multi-step in catena (chain-of-thought), capacità di riconoscere e correggere errori logici propri, generalizzazione a domini non rappresentati esplicitamente nel training. Queste capacità emergenti sono difficilmente spiegabili dal quadro del pappagallo stocastico — che implica una ricombinazione di pattern preesistenti — e suggeriscono che la distinzione tra "interpolazione sofisticata" e "ragionamento genuino" sia meno netta di quanto la metafora implichi. La questione rimane aperta, e questa apertura stessa mostra i limiti dell'immagine retorica come strumento analitico.

Il terzo problema concerne la natura del grounding. È corretto che i modelli linguistici non abbiano accesso diretto al mondo fisico e che le loro rappresentazioni siano mediate esclusivamente dalla struttura statistica del linguaggio. Ma questa obiezione è più complessa di quanto sembri. Il linguaggio non è una rappresentazione puramente arbitraria del mondo: è il prodotto di millenni di uso comunicativo da parte di agenti incorporati che hanno sviluppato sistemi linguistici in relazione alla loro esperienza fisica, sociale e cognitiva. Le regolarità semantiche e pragmatiche del linguaggio codificano, in forma compressa e indiretta, una quantità notevole di conoscenza sul mondo. Invocare l'assenza di grounding diretto come argomento conclusivo significa ignorare che il grounding è un continuum e non una proprietà binaria — la domanda pertinente non è "il sistema è grounded o no?" ma "quanto e in che modo?"

Il quarto problema è la dinamica empirica del campo. Alcune delle critiche più stringenti di LeCun agli LLM — la difficoltà nel ragionamento multi-step, la fragilità rispetto a variazioni superficiali degli input, l'incapacità di ragionare su situazioni fisiche nuove — si sono indebolite con l'introduzione del chain-of-thought prompting,

del reinforcement learning from human feedback, delle architetture multimodali che integrano percezione visiva e linguistica. Questo non refuta le critiche di principio, ma suggerisce che il confine tra "fluenza senza comprensione" e "comprensione" sia meno nitido di quanto la metafora del pappagallo suggerisca, e che la traiettoria del campo non sia necessariamente quella di un soffitto strutturale già raggiunto.

Queste considerazioni non invalidano la proposta di LeCun. L'intuizione che la comprensione richieda interazione con il mondo fisico, rappresentazioni causali e capacità di pianificazione è scientificamente fondata e sperimentalmente testabile. Esse suggeriscono però che la critica agli LLM, nella sua versione più forte e retorica, sia più problematica di quanto la sua presentazione sembri, e che la proposta di JEPa debba essere valutata per i propri meriti piuttosto che come semplice alternativa a sistemi ritenuti radicalmente inadeguati.

4. Il paper fondativo (2022): struttura e argomenti

"A Path Towards Autonomous Machine Intelligence" (LeCun, 2022) è un documento inconsueto nel panorama della ricerca contemporanea sull'intelligenza artificiale. Non presenta risultati sperimentali misurati su benchmark, non propone un modello addestrato su un dataset specifico, non ottimizza una funzione di perdita particolare. È invece un documento programmatico di largo respiro che combina argomenti di neuroscienza computazionale, psicologia cognitiva, teoria del controllo e apprendimento automatico in un quadro architeturale coerente. La sua struttura modulare — sei componenti principali con interfacce ben definite — è essa stessa una dichiarazione teorica: che l'intelligenza non sia una proprietà emergente di un unico sistema end-to-end, ma il risultato dell'integrazione di componenti specializzati con funzioni distinte, una posizione che richiama tanto la psicologia cognitiva modulare quanto le architetture di controllo della robotica.

Prima di analizzare i singoli componenti, è utile contestualizzare la proposta all'interno della tradizione dei sistemi di controllo ottimo e del reinforcement learning model-based. L'architettura lecuniana può essere vista come un tentativo di generalizzare la struttura di un controller ottimo in spazio di stato a un agente capace di apprendimento autonomo dal mondo fisico. La dinamica di un sistema di controllo classico è:

$$s_{t+1} = f(s_t, a_t)$$

dove $s_t \in \mathbb{R}^n$ è lo stato al tempo t , $a_t \in \mathbb{R}^m$ è l'azione, e $f: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ è la funzione di transizione di stato. Il controllore ottimo minimizza il costo accumulato:

$$J(a_{\{0:T\}}) = \sum_{t=0}^{T-1} c(s_t, a_t)$$

dove $c : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ è la funzione di costo istantaneo. Nella proposta di LeCun, f diventa il Modello di Mondo appreso, c diventa il modulo di costo apprendibile, e l'intero framework viene esteso per operare in spazio latente anziché in spazio di stato osservato — il che costituisce la trasformazione concettuale cruciale.

4.1 Il modulo percettivo e la gerarchia delle rappresentazioni

Il modulo percettivo ha il compito di trasformare gli input sensoriali grezzi — immagini, suoni, segnali propriocettivi e tattili — in rappresentazioni interne compatte e informative. La caratteristica fondamentale che LeCun richiede a queste rappresentazioni è la loro struttura gerarchica, composizionale e invariante rispetto alle trasformazioni irrilevanti per il compito.

Questo requisito può essere formalizzato attraverso il principio dell'Information Bottleneck (IB), introdotto da Tishby, Pereira e Bialek nel 2000. Data una variabile di input X , una variabile target Y (che in questo contesto può essere interpretata come la struttura causale latente del mondo), e una rappresentazione compressa $Z = f_\theta(X)$, il principio IB afferma che la rappresentazione ottimale massimizza l'informazione mutua $I(Z; Y)$ — quanto Z è informativa rispetto a ciò che conta — mentre minimizza $I(X; Z)$ — quanto della complessità di X viene ritenuta:

$$\mathcal{L}_{IB}(\theta) = I(Z; Y) - \beta \cdot I(X; Z)$$

dove $\beta \in \mathbb{R}_+$ è un parametro di Lagrange che controlla il trade-off tra fedeltà al target e compressione dell'input. La soluzione ottima di questo funzionale seleziona le rappresentazioni che preservano soltanto l'informazione di X causalmente rilevante per Y , comprimendo tutto ciò che è rumore, ridondanza o variabilità irrilevante.

L'informazione mutua in forma esplicita per variabili continue è:

$$I(Z; Y) = \iint p(z, y) \log [p(z, y) / (p(z) p(y))] dz dy$$

che misura la riduzione di entropia di Y data la conoscenza di Z (o equivalentemente la riduzione di entropia di Z data Y). In termini di entropia differenziale:

$$I(Z; Y) = H(Y) - H(Y | Z) = H(Z) - H(Z | Y)$$

La difficoltà computazionale di questo principio nel caso non supervisionato — in cui Y non è un'etichetta osservabile ma la struttura causale latente del mondo — è che l'ottimizzazione diretta di $I(Z; Y)$ è intrattabile. LeCun affronta questo problema attraverso un proxy autosupervisionato: le rappresentazioni apprese da JEPA cercano di predire porzioni future o mascherate dell'input, nella speranza che questo compito di predizione funzioni come proxy sufficientemente buono della struttura causale. La qualità di questo proxy è una delle questioni aperte centrali della proposta.

Il modulo percettivo è concepito per operare su input multimodali — visivi, uditivi, propriocettivi — producendo rappresentazioni in spazi latenti separati che vengono poi integrati da meccanismi di fusione cross-modale. Questa struttura multimodale è motivata dalla considerazione che la comprensione del mondo fisico richiede l'integrazione di informazioni provenienti da canali sensoriali diversi: la stessa scena fisica viene percepita visivamente, udibilmente (i suoni prodotti dagli impatti tra oggetti) e tattilmente (per un agente robotico), e la convergenza di queste informazioni produce una rappresentazione più robusta di quanto possano offrire i singoli canali isolati.

4.2 Il Modello di Mondo: struttura formale e proprietà richieste

Il Modello di Mondo è il nucleo della proposta lecuniana e il componente che la distingue più nettamente dagli approcci dominanti. La sua funzione è duplice: predire le conseguenze delle azioni sull'ambiente e rappresentare lo stato del mondo in modo sufficientemente ricco da supportare la pianificazione a diversi livelli di astrazione temporale.

Formalmente, un Modello di Mondo è una funzione che mappa la coppia (stato latente corrente s_t , azione a_t , variabili latenti stocastiche z_t) a una distribuzione sugli stati latenti futuri. Nella formulazione di LeCun:

$$\hat{s}_{t+1} = M_\theta(s_t, a_t, z_t)$$

dove $s_t \in \mathbb{R}^{n_s}$ è la rappresentazione dello stato corrente nello spazio latente, $a_t \in \mathbb{R}^{n_a}$ è la rappresentazione dell'azione, $z_t \in \mathbb{R}^{n_z}$ è un vettore di variabili latenti che codifica l'incertezza irriducibile del mondo (aspetti dello stato futuro che non sono determinati dallo stato e dall'azione correnti), e M_θ è la funzione di transizione parametrizzata. La formulazione non è probabilistica nel senso classico di stimare una densità esplicita su s_{t+1} : è invece deterministica condizionatamente a z_t , il che sposta la modellazione dell'incertezza sulla distribuzione di z_t anziché sull'output del modello.

Questa scelta è motivata da un'osservazione fondamentale: nel mondo fisico reale, l'incertezza sullo stato futuro proviene in parte da rumore genuinamente aleatorio (il rimbalzo caotico di un liquido, le fluttuazioni termiche) e in parte da incomplete informazioni sullo stato attuale (aspetti del mondo non osservati dalla posizione corrente). Un modello che tenta di catturare questa incertezza come distribuzione sulla sua uscita deve affrontare il problema dell'apprendimento di distribuzioni multimodali complesse nello spazio degli stati — un compito notoriamente difficile per i modelli neurali. Spostare l'incertezza in z_t esplicito permette al modello di rimanere deterministico condizionatamente e di delegare la modellazione dell'incertezza a un componente separato.

La distinzione tra predizione in spazio latente e predizione in spazio osservativo è cruciale per comprendere la proposta. Un modello di transizione classico, come i

modelli generativi di video o i modelli di dinamica usati nel reinforcement learning model-based, stima:

$$P(x_{t+1} \mid x_t, a_t)$$

dove x_t è l'osservazione grezza (per esempio, i pixel di un frame video). Questa formulazione impone al modello di predire tutti i dettagli dell'osservazione futura — texture, illuminazione, rumore sensoriale, fluttuazioni caotiche — anche quelli irrilevanti per qualsiasi compito di pianificazione. Il costo di questa predizione è triplice: computazionale (la dimensionalità di x_{t+1} può essere dell'ordine di 10^6 per immagini ad alta risoluzione), statistica (il modello deve imparare a generare distribuzioni complesse su insiemi di pixel), e epistemica (l'addestramento può essere dominato dalla modellazione di dettagli irrilevanti a scapito della cattura di strutture causali di alto livello).

La predizione in spazio latente proposta da LeCun stima invece:

$$\hat{s}_{t+1} = M_\theta(s_t, a_t, z_t), \text{ con } s_t = E_\theta(x_t)$$

dove E_θ è l'encoder che mappa le osservazioni grezze nello spazio latente. Il modello predice direttamente la rappresentazione latente dello stato futuro, non lo stato osservativo grezzo. Questo significa che le variazioni dell'osservazione grezza che non alterano la rappresentazione latente — e che sono quindi irrilevanti per i compiti che la rappresentazione latente è progettata per supportare — non contribuiscono alla perdita di addestramento e non devono essere modellate.

Il Modello di Mondo deve essere capace di operare a livelli di astrazione multipli: deve predire sia le proprietà fisiche immediate (la posizione e la velocità di un oggetto nel frame successivo) sia le strutture di alto livello a lungo termine (questo oggetto colliderà con il muro tra tre secondi). Questa gerarchia temporale è implementata nella proposta di LeCun attraverso l'architettura gerarchica del pianificatore, che decompone obiettivi a lungo termine in sotto-obiettivi progressivamente più concreti e utilizza Modelli di Mondo a diverse scale temporali.

4.3 Il modulo di memoria: architettura e principi formali

Il modulo di memoria nell'architettura lecuniana ha una struttura duale che distingue tra memoria di lavoro a breve termine e memoria episodica a lungo termine. Questa distinzione rispecchia la bipartizione classica delle neuroscienze cognitive tra working memory — il sistema che mantiene attive le informazioni rilevanti per il compito corrente — e memoria episodica — il sistema che immagazzina e recupera eventi passati strutturati spaziotemporalmente.

La memoria di lavoro a breve termine mantiene una rappresentazione compressa del contesto corrente: lo stato latente percepito s_t , la sequenza delle ultime azioni eseguite $a_{t-k:t}$, le predizioni generate dal Modello di Mondo per il futuro immediato, e gli obiettivi correnti del pianificatore. Formalmente, questa può essere modellata come uno stato ricorrente che viene aggiornato a ogni time step:

$$c_t = g_\theta(c_{t-1}, s_t, a_{t-1})$$

dove g_θ è una funzione di aggiornamento del contesto parametrizzata — strutturalmente analoga a una cella LSTM o GRU, ma operante nello spazio latente delle rappresentazioni del Modello di Mondo anziché in quello dei token linguistici.

La memoria a lungo termine è concepita come una struttura content-addressable che immagazzina episodi passati recuperabili per somiglianza con la query corrente. LeCun propone di implementare questo attraverso reti di Hopfield moderne — nella formulazione di Ramsauer et al. (2020) discussa nella sezione precedente — che possono memorizzare e recuperare un numero esponenzialmente grande di pattern attraverso aggiornamenti per gradient descent sull'energia. Il recupero da questa memoria avviene attraverso:

$$m_t = \text{softmax}(\beta K^T s_t) \cdot V$$

dove $K \in \mathbb{R}^{(M \times d)}$ è la matrice delle chiavi (rappresentazioni latenti degli episodi memorizzati), $V \in \mathbb{R}^{(M \times d)}$ è la matrice dei valori (contenuti degli episodi), s_t è la query corrente, e β è un parametro di temperatura. Questa operazione ha la stessa forma algebrica dell'attenzione Transformer, ma il suo ruolo funzionale è diverso: non è un meccanismo di elaborazione del contesto all'interno di una sequenza, ma un meccanismo di recupero episodico da una memoria esterna che viene aggiornata incrementalmente durante l'esperienza del sistema.

L'integrazione della memoria a lungo termine con il Modello di Mondo è fondamentale per la generalizzazione per analogia: il sistema dovrebbe essere capace di recuperare episodi passati strutturalmente analoghi alla situazione corrente e di sfruttarne le rappresentazioni per migliorare le predizioni del Modello di Mondo in situazioni nuove. Questo meccanismo è il corrispondente computazionale dell'apprendimento one-shot e few-shot che gli esseri umani mostrano in compiti di ragionamento fisico: avendo visto una volta cosa succede quando si spinge un oggetto vicino al bordo di un tavolo, un essere umano sa immediatamente cosa aspettarsi in situazioni simili, anche se la geometria esatta della situazione non è stata vista prima.

4.4 Il modulo di costo e la funzione energetica: fondamenti formali degli EBM

Il modulo di costo è, concettualmente, la sede dei valori del sistema. La sua funzione è assegnare un valore scalare — costo o energia — a ogni coppia (stato, azione) o a ogni traiettoria, guidando la pianificazione verso stati preferibili. Questa architettura adotta la tradizione dei modelli energetici (Energy-Based Models, EBM) che LeCun ha sviluppato e difeso nel corso di vent'anni di ricerca, presentandola come alternativa al paradigma probabilistico normalizzato.

Un EBM assegna a ogni configurazione della variabile (x, y) un valore di energia scalare $E_\theta(x, y) \in \mathbb{R}$, con la convenienza che configurazioni

compatibili — paia (x, y) che corrispondono a relazioni reali nel dominio — abbiano energia bassa, e configurazioni incompatibili abbiano energia alta. La distribuzione di Gibbs associata è:

$$P_{\theta}(y | x) = \exp(-E_{\theta}(x, y)) / Z_{\theta}(x)$$

dove $Z_{\theta}(x) = \int \exp(-E_{\theta}(x, y)) dy$ è la funzione di partizione — la costante di normalizzazione che rende $P_{\theta}(y|x)$ una distribuzione di probabilità valida. Il problema fondamentale dei modelli probabilistici normalizzati è che $Z_{\theta}(x)$ è in generale intrattabile per variabili continue di alta dimensionalità: il calcolo dell'integrale richiede o stime Monte Carlo molto costose, o approssimazioni variazionali che introducono bias.

Il punto di forza degli EBM è che non richiedono la normalizzazione. L'addestramento non mira a stimare $Z_{\theta}(x)$ ma a fare in modo che E_{θ} sia bassa per le configurazioni reali e alta per quelle spurie. Questo si ottiene attraverso diverse strategie di addestramento:

La contrastive divergence (CD), introdotta da Hinton (2002), aggiorna i parametri attraverso:

$$\partial E_{\theta} / \partial \theta_{CD} = \mathbb{E}_{\{(x,y) \sim P_{data}\}} [\partial E_{\theta}(x,y) / \partial \theta] - \mathbb{E}_{\{(x,\tilde{y}) \sim P_{\theta}\}} [\partial E_{\theta}(x,\tilde{y}) / \partial \theta]$$

Il primo termine spinge verso il basso l'energia delle configurazioni reali ("dati positivi"); il secondo spinge verso l'alto l'energia di configurazioni campionate dalla distribuzione del modello ("dati negativi"). Il campionamento dal secondo termine richiede MCMC (tipicamente Gibbs sampling o Langevin dynamics), che è computazionalmente costoso e può essere affetto da problemi di mixing lento.

Il margin loss (o hinge loss per gli EBM), nella formulazione di LeCun et al. (2006):

$$\mathcal{L}_{margin}(\theta) = \sum_{\{(x,y) \in \mathcal{D}\}} \sum_{\{\tilde{y} \neq y\}} \max(0, m - E_{\theta}(x, \tilde{y}) + E_{\theta}(x, y))$$

spinge l'energia dei dati positivi al di sotto di quella dei negativi di un margine m , senza richiedere la normalizzazione esplicita. Qui $\tilde{y}$ sono "dati negativi" campionati dallo spazio delle configurazioni, tipicamente attraverso perturbazione delle variabili di risposta y .

Nel contesto della pianificazione lecuniana, il modulo di costo $C_{\theta}(s, a)$ svolge la funzione dell'energia: assegna un valore scalare a ogni coppia (stato latente, azione), con la proprietà che stati e azioni corrispondenti a comportamenti desiderabili abbiano costo basso e quelli indesiderabili abbiano costo alto. Il costo totale di una traiettoria $(s_0, a_0, s_1, a_1, \dots, s_T)$ è:

$$J(a_{0:T}) = \sum_{t=0}^{T-1} C_{\theta}(s_t, a_t) + C_T(s_T)$$

dove C_T è un costo terminale che misura la distanza dello stato finale dall'obiettivo. La pianificazione ottimale consiste nel trovare la sequenza di azioni che minimizza J , vincolata alla dinamica del Modello di Mondo:

$$a^*\{0:T\} = \underset{a_{\{0:T\}}}{\operatorname{argmin}} J(a_{\{0:T\}}) \text{ s.t. } s_{t+1} = M_\theta(s_t, a_t, z_t)$$

Il modulo di costo tiene traccia di due tipologie di obiettivi. Gli obiettivi intrinseci (hardwired) sono componenti del costo progettati dall'ingegnere e non modificabili attraverso l'apprendimento: ad esempio, vincoli sulla sicurezza fisica dell'agente (evitare stati che causano danni al hardware), vincoli sull'uso delle risorse energetiche, o obiettivi di riduzione dell'incertezza sulla rappresentazione del mondo. Gli obiettivi acquisiti (trainable) sono componenti del costo che il sistema impara attraverso l'esperienza, l'imitazione di agenti esperti, o il feedback da supervisori umani. Questa distinzione riflette la dualità biologica tra sistemi di valore innati (i meccanismi omeostatici, l'avversione al dolore fisico) e sistemi di valore appresi (le preferenze culturali, gli obiettivi socialmente mediati).

4.5 Il pianificatore gerarchico: ottimizzazione in spazio latente

Il pianificatore gerarchico è il componente responsabile della generazione di sequenze di azioni ottimali rispetto al modulo di costo, sfruttando le predizioni del Modello di Mondo. La sua struttura gerarchica riflette la tesi che la pianificazione efficace richieda la decomposizione di obiettivi ad alto livello in sotto-obiettivi progressivamente più concreti — un principio che compare sia nelle teorie psicologiche della pianificazione umana (la hierarchical task analysis di Annett, 1967) sia nelle architetture di controllo della robotica.

Il pianificatore opera come un ottimizzatore che esegue gradient descent nello spazio delle azioni (o nel spazio latente delle azioni) rispetto al costo accumulato predetto dal Modello di Mondo. Formalmente, dato uno stato iniziale s_0 e un obiettivo s_{goal} , il pianificatore risolve il problema di ottimizzazione:

$$a^*\{0:T\} = \underset{a_{\{0:T\}}}{\operatorname{argmin}} [\sum_{t=0}^T C_\theta(s_t, a_t) + C_T(s_T, s_{\text{goal}})]$$

con il vincolo di dinamica $s_{t+1} = M_\theta(s_t, a_t, z_t)$. Poiché M_θ e C_θ sono funzioni differenziabili parametrizzate da reti neurali, il gradiente dell'obiettivo rispetto alle azioni può essere calcolato attraverso backpropagation nel tempo (BPTT) attraverso il Modello di Mondo:

$$\partial J / \partial a_t = \partial C_\theta(s_t, a_t) / \partial a_t + (\partial s_{t+1} / \partial a_t)^T \cdot (\partial J / \partial s_{t+1})$$

dove $\partial s_{t+1} / \partial a_t = \partial M_\theta(s_t, a_t, z_t) / \partial a_t$ è il jacobiano del Modello di Mondo rispetto all'azione. Il pianificatore aggiorna le azioni attraverso gradient descent:

$$a_t \leftarrow a_t - \eta \cdot \partial J / \partial a_t$$

iterando fino alla convergenza. Questo approccio — noto come pianificazione model-based attraverso backpropagation — non è nuovo: compare nei lavori di Werbos (1990) sulla backpropagation through time, di Dayan e Hinton (1993) sulle reti feudali (feudal nets), e di Åström e collaboratori nella teoria del controllo ottimo. LeCun lo combina con la predizione in spazio latente e l'architettura gerarchica in un modo che cerca di superare alcune delle limitazioni note, in particolare il problema del gradiente che svanisce attraverso molti passi di predizione e l'instabilità numerica dell'ottimizzazione in spazio di azione continuo.

La struttura gerarchica introduce livelli multipli di pianificazione. Al livello più alto, il pianificatore produce sotto-obiettivi astratti s_goal^H nello spazio latente di alto livello, con un orizzonte temporale lungo. Al livello intermedio, ogni sotto-obiettivo astratto viene tradotto in sotto-obiettivi più concreti s_goal^{H-1} con orizzonte più breve. Al livello più basso, i sotto-obiettivi concreti vengono tradotti in sequenze di azioni primitive eseguibili dal modulo attoriale. Ogni livello utilizza il proprio Modello di Mondo, operante alla granularità temporale appropriata: il livello alto può avere un passo temporale dell'ordine dei secondi, il livello basso dell'ordine dei millisecondi.

Formalmente, la pianificazione al livello h è:

$$a^{*h}\{0:T_h\} = \underset{a^h\{0:T_h\}}{\operatorname{argmin}} \sum_{t=0}^{T_h} C_{\theta^h}(s^h_t, a^h_t)$$

$$\text{s.t. } s^{h}_{t+1} = M_{\theta^h}(s^h_t, a^h_t, z^h_t)$$

dove le "azioni" al livello h sono i sotto-obiettivi prodotti dalla pianificazione al livello h per il livello $h-1$. Questa struttura è analoga alle reti feudali di Dayan e Hinton (1993), all'opzioni framework di Sutton, Precup e Singh (1999) nel reinforcement learning gerarchico, e alle architetture MAXQ di Dietterich (2000), ma si distingue da esse per operare interamente in spazio latente continuo anziché su spazio di stati e azioni discreti.

4.6 Il modulo attoriale

Il modulo attoriale traduce i piani generati dal pianificatore in comandi motori eseguibili sull'ambiente. La sua funzione è di controllo a basso livello: ricevuti i sotto-obiettivi concreti dal livello più basso del pianificatore, genera sequenze di comandi (forze, velocità, posizioni di giuntura per un robot) che realizzano quei sotto-obiettivi.

La distinzione tra pianificatore e attore rispecchia una bipartizione teoricamente importante nelle neuroscienze cognitive: quella tra il sistema "cosa fare" — implementato prevalentemente nelle aree prefrontali e nelle strutture dorsali — e il sistema "come farlo" — implementato prevalentemente nelle aree parietali e nei circuiti motori. Questa doppia dissociazione è documentata empiricamente sia nei dati di neuroimaging (studi fMRI e EEG su compiti di pianificazione motoria) sia nella clinica neurologica (la distinzione tra aprassia ideativa — incapacità di pianificare la

sequenza delle azioni — e aprassia ideomotoria — incapacità di eseguirle motorialmente pur mantenendo la pianificazione integra). La sua incorporazione nell'architettura lecuniana è uno degli esempi più espliciti di come la proposta aspiri a una corrispondenza funzionale con le architetture cognitive biologiche.

5. JEPA: Joint Embedding Predictive Architecture

5.1 Il principio fondamentale: predizione in spazio latente

JEPA — Joint Embedding Predictive Architecture — è la proposta architettonica concreta che implementa i principi del Modello di Mondo nell'ambito dell'apprendimento autosupervisionato. Il suo principio organizzativo è la predizione in spazio latente: invece di apprendere a predire la rappresentazione osservativa grezza futura di un input (i pixel del frame successivo di un video, i token successivi in una sequenza), JEPA apprende a predire la rappresentazione latente dell'input futuro — ovvero predice direttamente nello spazio delle rappresentazioni compresse anziché nello spazio delle osservazioni.

La motivazione formale di questa scelta può essere articolata attraverso il principio di minima descrizione. Sia x la variabile di input osservativa (un'immagine, un frame video) e sia $z = E_\theta(x)$ la sua rappresentazione latente. Un buon encoder E_θ dovrebbe soddisfare:

$$H(z) \ll H(x)$$

ovvero la rappresentazione latente deve avere entropia molto inferiore all'input grezzo — deve comprimere l'informazione irrilevante. Se l'encoder è addestrato a catturare le strutture causalmente rilevanti, la predizione nel spazio z ha un compito molto più facile di una predizione nel spazio x : nel mondo fisico, molti aspetti dello stato osservativo futuro (texture, illuminazione, rumore) sono non-causalmente correlati con lo stato corrente — non sono predicibili dallo stato e dall'azione ma dipendono da fattori stocastici irrilevanti. Predire in spazio latente significa predire soltanto le proprietà strutturali predicibili, ignorando le fluttuazioni non predicibili che dominano la varianza nel spazio osservativo.

5.2 Struttura formale dell'architettura JEPA

L'architettura JEPA si compone di tre reti neurali principali che interagiscono attraverso un protocollo di addestramento autosupervisionato.

Il primo componente è l'encoder del contesto $E_x : \mathcal{X} \rightarrow \mathbb{R}^d$. Dato un input di contesto x (una porzione parziale dell'immagine, una finestra temporale di un video, l'audio precedente), E_x produce una rappresentazione latente:

$$s_x = E_x(x) \in \mathbb{R}^d$$

Il secondo componente è l'encoder del target $E_y : \mathcal{X} \rightarrow \mathbb{R}^d$. Dato l'input target y (la porzione mascherata dell'immagine, il frame futuro del video, il segmento audio successivo), E_y produce la rappresentazione target:

$$s_y = E_y(y) \in \mathbb{R}^d$$

Il terzo componente è il predittore $P : \mathbb{R}^d \times \mathbb{R}^{(d_z)} \rightarrow \mathbb{R}^d$. Dati la rappresentazione del contesto s_x e un vettore di variabili latenti z che codifica l'incertezza irriducibile, il predittore genera una predizione della rappresentazione target:

$$\hat{s}_y = P(s_x, z) \in \mathbb{R}^d$$

L'obiettivo di addestramento è minimizzare la distanza (in norma L2, o in qualche metrica appropriata sullo spazio latente) tra la predizione $\hat{s}_y$ e il target effettivo s_y :

$$\mathcal{L}_{\text{JEPA}}(\theta_x, \theta_y, \theta_P) = \mathbb{E}_{\{(x,y) \sim \mathcal{D}\}} [\|P(E_x(x), z) - E_y(y)\|^2]$$

dove l'aspettazione è calcolata sulla distribuzione congiunta delle coppie contesto-target nel corpus di addestramento.

Il problema critico di questa formulazione — comune a tutti i metodi autosupervisionati basati su predizione in spazio latente — è il collasso rappresentazionale. Se l'encoder è libero di apprendere qualunque funzione, la soluzione triviale che minimizza $\mathcal{L}_{\text{JEPA}}$ è quella che collassa tutte le rappresentazioni in un singolo punto: $E_x(x) = c$ e $E_y(y) = c$ per ogni x, y , dove c è una costante. In questo caso $\|P(c, z) - c\|^2 = 0$ per il predittore banale $P(c, z) = c$, ma le rappresentazioni sono prive di qualunque contenuto informativo.

Questo problema del collasso ha soluzioni diverse in JEPA rispetto ai metodi contrastivi:

Nei metodi contrastivi (SimCLR, MoCo, BYOL, SimSiam), il collasso viene prevenuto attraverso un termine di distanziamento che spinge le rappresentazioni di campioni diversi a essere distanti:

$$\mathcal{L}_{\text{contrastive}} = - \sum_i \log \left[\frac{\exp(s_i \cdot s_i^+ / \tau)}{\sum_{j \neq i} \exp(s_i \cdot s_j / \tau)} \right]$$

dove s_i^+ è la rappresentazione di una vista augmented dello stesso campione x_i , s_j sono le rappresentazioni di campioni negativi (campioni diversi), e τ è la temperatura. Questa perdita massimizza la somiglianza tra viste diverse dello stesso campione (positive pairs) e minimizza la somiglianza tra campioni diversi (negative pairs), prevenendo il collasso per costruzione.

In JEPA, il collasso viene prevenuto attraverso una combinazione di tecniche architetturali. La tecnica principale, usata in I-JEPA, è il momentum encoder per l'encoder del target: E_y non è addestrato direttamente attraverso backpropagation

della perdita, ma è aggiornato come media esponenziale mobile dell'encoder del contesto E_x :

$$\theta_y \leftarrow \tau \cdot \theta_y + (1 - \tau) \cdot \theta_x$$

con $\tau \in [0.99, 0.999]$ tipicamente. Questa tecnica, usata originalmente in MoCo (He et al., 2020) e BYOL (Grill et al., 2020), introduce una forma di "stop gradient" implicita che rompe la simmetria del collasso: l'encoder del target "segue" l'encoder del contesto con un ritardo esponenzialmente pesato, creando un target moving che il predittore deve inseguire — un obiettivo non banale da raggiungere con il collasso.

Un secondo meccanismo anti-collasso è la scelta della struttura di mascheramento: in I-JEPA, le patch mascherate (target) sono scelte in modo da coprire regioni sufficientemente grandi e semanticamente ricche dell'immagine, garantendo che la predizione richieda strutture rappresentazionali genuine piuttosto che semplici interpolazioni locali.

5.3 I-JEPA: implementazione su immagini e analisi formale

I-JEPA (Image-JEPA) è la prima implementazione concreta dell'architettura, presentata da Assran e collaboratori (2023) nel paper "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture", accettato a CVPR 2023. Il sistema è addestrato su ImageNet-1K attraverso un protocollo di masked patch prediction.

L'input è un'immagine $I \in \mathbb{R}^{(H \times W \times 3)}$ che viene suddivisa in patch non-sovrapponibili di dimensione $p \times p$, producendo una sequenza di $N = (H/p)(W/p)$ patch. Ogni patch viene proiettata in un vettore d -dimensionale attraverso un'operazione lineare (patch embedding). Le patch vengono divise in due insiemi: il contesto C (patch visibili, usate come input all'encoder del contesto) e il target T (patch mascherate, la cui rappresentazione latente deve essere predetta). Il protocollo di mascheramento in I-JEPA usa blocchi contigui piuttosto che patch casuali isolate, e utilizza target multipli (tipicamente 4) per ogni contesto:

$$x_C = \{\text{patch}_i : i \in C\}, y^{(k)}_T = \{\text{patch}_j : j \in T^{(k)}\}, k = 1, \dots, K$$

L'encoder del contesto è un Vision Transformer (ViT) che elabora soltanto le patch del contesto, producendo:

$$s_C = E_x(x_C) \in \mathbb{R}^{|C| \times d}$$

Il predittore, un Transformer più leggero, riceve s_C e i token di posizione delle patch target, producendo le predizioni:

$$\hat{s}^{(k)}_T = P(s_C, \text{pos}^{(k)}_T) \in \mathbb{R}^{|T^{(k)}| \times d}$$

L'encoder del target (con momentum) elabora l'intera immagine e produce le rappresentazioni target:

$$s^{(k)}_T = E_y(y^{(k)}_T) \in \mathbb{R}^{|T^{(k)}| \times d}$$

La perdita di addestramento è:

$$\mathcal{L}_{\{I\text{-JEPA}\}} = (1/K) \sum_{k=1}^K (1/|T^{(k)}|) \sum_{j \in T^{(k)}} \| \hat{s}^{(k)}_{\{T,j\}} - \text{sg}(s^{(k)}_{\{T,j\}}) \|^2$$

dove $\text{sg}(\cdot)$ indica lo stop-gradient — l'operazione che blocca il gradiente dall'encoder del target (il quale viene aggiornato solo attraverso il momentum, non attraverso la backpropagation della perdita). Il fattore di normalizzazione garantisce che la perdita sia comparabile indipendentemente dalla dimensione del target.

I risultati di I-JEPA su ImageNet-1K in configurazione di linear probing — in cui le rappresentazioni apprese vengono valutate attaccando un classificatore lineare senza fine-tuning del backbone — mostrano una qualità competitiva con i migliori metodi autosupervisionati. Più interessante per la valutazione della proposta di LeCun è la struttura qualitativa delle rappresentazioni: analisi di nearest-neighbor sullo spazio latente mostrano che le rappresentazioni di I-JEPA catturano similarità semantiche (oggetti della stessa categoria sono vicini nello spazio latente) con meno dipendenza dalle caratteristiche texturali rispetto a metodi come MAE (He et al., 2022) che prevedono direttamente i pixel. Questo è interpretato come evidenza che la predizione in spazio latente produce rappresentazioni più "concettuali" e meno "sensoriali".

5.4 V-JEPA: estensione alla modalità video

V-JEPA (Video-JEPA) estende l'architettura alla modalità video, presentato da Bardes e collaboratori (2024) nel paper "V-JEPA: Latent Video Prediction for Visual Representation Learning". L'estensione introduce sfide considerevolmente più complesse rispetto al caso delle immagini statiche: il segnale video contiene informazioni causali e temporali che il sistema deve apprendere a sfruttare, e la dimensionalità dell'input è ordini di grandezza superiore.

Formalmente, un video V è una sequenza di T frame: $V = (I_1, I_2, \dots, I_T)$, con $I_t \in \mathbb{R}^{(H \times W \times 3)}$. Il protocollo di mascheramento spazio-temporale di V-JEPA estende il mascheramento per blocchi di I-JEPA alla dimensione temporale: vengono mascherati volumi spazio-temporali contigui (blocchi di patch consecutive sia nello spazio che nel tempo). L'encoder del contesto elabora le patch visibili prodotte dopo il mascheramento:

$$s_C = E_x(\{(patch_{\{i,t\}}) : (i,t) \notin M\})$$

dove M è il set di posizioni mascherate nello spazio-tempo. Il predittore deve poi ricostruire le rappresentazioni latenti delle patch mascherate, condizionatamente al contesto:

$$\hat{s}_M = P(s_C, pos_M) \in \mathbb{R}^{|M| \times d}$$

La perdita è analoga a quella di I-JEPA, con target dati dall'encoder con momentum che elabora le patch mascherate.

L'architettura di V-JEPA usa un Video Vision Transformer (ViViT) come backbone, che applica self-attention sia spazialmente (all'interno di ogni frame) sia temporalmente (tra frame corrispondenti in tempi diversi). La fattorizzazione dell'attenzione spazio-temporale permette di gestire le sequenze video senza la complessità quadratica in $T \cdot N$ che si avrebbe con l'attenzione piena.

I risultati empirici di V-JEPA mostrano prestazioni competitive su benchmark di comprensione di azioni (Something-Something V2, Kinetics-400) e di ragionamento fisico semplice. Particolarmente notevole è la capacità di V-JEPA di generalizzare a variazioni di prospettiva e illuminazione che non erano rappresentate nel training — una proprietà attesa se le rappresentazioni latenti catturano strutture invarianti anziché pattern sensoriali superficiali. LeCun e collaboratori interpretano questi risultati come conferma del principio che la predizione in spazio latente produce rappresentazioni qualitativamente superiori rispetto ai metodi che prevedono direttamente i pixel.

5.5 Confronto formale con approcci contrastivi e generativi

Il posizionamento di JEPA rispetto agli altri approcci di apprendimento autosupervisionato può essere chiarito attraverso un'analisi formale delle loro funzioni obiettivo e delle proprietà delle rappresentazioni che producono.

Sia $f : \mathcal{X} \rightarrow \mathbb{R}^d$ un encoder e sia $\mathcal{T}$ un insieme di trasformazioni (augmentations) che preservano l'identità semantica del campione. I metodi contrastivi come SimCLR (Chen et al., 2020) massimizzano la concordanza tra viste diverse dello stesso campione minimizzando contemporaneamente la concordanza tra campioni diversi. La funzione obiettivo NT-Xent di SimCLR è:

$$\mathcal{L}_{\text{SimCLR}} = - (1/N) \sum_{i=1}^N \log \left[\frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_{j=1, j \neq i}^{2N} \exp(\text{sim}(z_i, z_j)/\tau)} \right]$$

dove $z_i = g(f(T(x_i)))$, T è una trasformazione campionata uniformemente da $\mathcal{T}$, g è un proiettore MLP, e $\text{sim}(u, v) = u^T v / (\|u\| \|v\|)$ è la similarità coseno. Il denominatore somma su tutti i $2N - 1$ campioni nel batch tranne z_i stesso, usando i $2N - 2$ campioni negativi (campioni da $x_{\{j \neq i\}}$) più il campione positivo z_i^+ (l'altra augmentation dello stesso x_i).

Il limite principale dei metodi contrastivi è la dipendenza dalla scelta delle trasformazioni in $\mathcal{T}$. Le augmentations devono essere progettate manualmente per ciascuna modalità di input (e la scelta ottimale differisce per immagini, audio, testo, video), devono preservare la semantica dell'input ma variare le caratteristiche superficiali, e la loro scelta influenza in modo critico la qualità delle rappresentazioni apprese — le caratteristiche che le augmentations non alterano vengono apprese come invarianti, il che può essere desiderabile (l'illuminazione non cambia l'identità di un

oggetto) o indesiderabile (la pose non dovrebbe essere un'invarianza per compiti di riconoscimento tridimensionale).

I metodi generativi come il Masked Autoencoder (MAE, He et al., 2022) apprendono rappresentazioni attraverso la ricostruzione dell'input mascherato:

$$\mathcal{L}_{\text{MAE}} = (1/|M|) \sum_{\{i \in M\}} \|\hat{x}_i - x_i\|^2$$

dove $\hat{x}_i$ è la ricostruzione del pixel i nella posizione mascherata. MAE ha dimostrato prestazioni eccellenti su benchmark di classificazione e segmentazione, con un'implementazione semplice ed efficiente. Il suo limite principale, nella prospettiva lecuniana, è che la perdita di ricostruzione pixel alloca capacità rappresentazionale a dettagli texturali di alta frequenza che sono irrilevanti per la maggior parte dei compiti cognitivi — e che, crucialmente, non sono prevedibili a partire dalla struttura semantica dell'immagine ma dipendono da fluttuazioni stocastiche dell'acquisizione sensoriale.

JEPA può essere posizionato nel triangolo tra questi approcci: come i metodi contrastivi, non richiede la ricostruzione dello spazio osservativo e quindi non alloca capacità a dettagli irrilevanti; come i metodi generativi, modella la struttura predittiva dell'input attraverso una perdita di predizione anziché attraverso termini di distanziamento. La predizione in spazio latente è il meccanismo che rende possibile questo bilanciamento: si predice la struttura dell'input futuro (come nei generativi) ma nello spazio compresso che è stato progettato per non contenere i dettagli irrilevanti (come nei contrastivi, che ignorano le variazioni irrilevanti attraverso augmentation).

Il confronto può essere reso preciso attraverso l'analisi delle proprietà di equivarianza e invarianza delle rappresentazioni. Sia T_g una trasformazione parametrizzata da g (una rotazione di angolo θ , una traslazione di vettore v , ecc.). Si dice che una rappresentazione $f(x)$ è invariante rispetto a T_g se:

$$f(T_g(x)) = f(x) \text{ per ogni } x \text{ e ogni } g$$

Si dice equivariante rispetto a T_g se esiste una mappa q tale che:

$$f(T_g(x)) = q(g) \cdot f(x) \text{ per ogni } x \text{ e ogni } g$$

I metodi contrastivi producono rappresentazioni invarianti rispetto alle trasformazioni in $\mathcal{T}$ (per costruzione della funzione obiettivo), il che può essere eccessivamente distruttivo per compiti che richiedono la comprensione della struttura spaziale o temporale. JEPA non impone invarianza per costruzione, ma apprende le invarianze attraverso il compito di predizione: se due stati dell'input che differiscono per caratteristiche irrilevanti hanno la stessa evoluzione futura (stessa struttura latente futura), l'encoder apprende a rappresentarli nello stesso punto dello spazio latente. Questo approccio alla costruzione delle invarianze è più flessibile e adattivo rispetto alla specifica manuale delle augmentations nei metodi contrastivi.

5.6 Il problema del collasso e le strategie di prevenzione: un'analisi rigorosa

Il problema del collasso è la sfida tecnica centrale di tutti i metodi di apprendimento autosupervisionato basati su predizione o concordanza in spazio latente. Una comprensione rigorosa delle strategie di prevenzione del collasso usate da JEPA richiede un'analisi delle condizioni sotto cui il collasso avviene e delle proprietà che le soluzioni non-collassate devono soddisfare.

Il collasso si produce quando l'encoder mappa tutti gli input in un unico punto (collasso completo) o in una varietà di dimensione molto inferiore alla dimensione dello spazio latente (collasso parziale o dimensional collapse). Formalmente, il collasso completo della rappresentazione $z = E_\theta(x)$ avviene quando:

$$\text{Cov}(z) = \mathbb{E}[(z - \mathbb{E}[z])(z - \mathbb{E}[z])^T] \approx 0$$

ovvero quando la matrice di covarianza delle rappresentazioni è quasi singolare — tutte le direzioni dello spazio latente hanno varianza quasi nulla. Il collasso parziale è più insidioso: alcune direzioni hanno varianza non nulla (quelle catturate dal modello) ma molte altre hanno varianza nulla (direzioni non usate), riducendo effettivamente la dimensione utile della rappresentazione.

Diverse strategie contrastano questi fenomeni. Il momento esponenziale (EMA) usato in I-JEPA introduce un'asimmetria tra l'encoder del contesto E_x (aggiornato attraverso gradient descent sulla perdita) e l'encoder del target E_y (aggiornato attraverso EMA di E_x). Questa asimmetria rompe la simmetria del problema di ottimizzazione che renderebbe triviale il collasso: se entrambi gli encoder fossero identici e aggiornati dalla stessa perdita, la soluzione ottima sarebbe il collasso; con l'EMA, l'encoder del target è un "oracolo ritardato" che il predittore deve seguire, e la perdita non può essere minimizzata collassando entrambi gli encoder nello stesso punto perché il target cambia con un ritardo che dipende da τ .

Il metodo VICReg (Bardes, Ponce e LeCun, 2021) propone un'alternativa esplicita al collasso basata su tre termini di regolarizzazione che possono essere applicati anche ai metodi JEPA:

$$\mathcal{L}_{\text{VICReg}} = \lambda \cdot \mathcal{L}_{\text{var}} + \mu \cdot \mathcal{L}_{\text{inv}} + \nu \cdot \mathcal{L}_{\text{cov}}$$

Il termine di invarianza $\mathcal{L}_{\text{inv}} = \|z - z'\|^2$ avvicina le rappresentazioni di due viste dello stesso campione. Il termine di varianza $\mathcal{L}_{\text{var}} = \sum_j \max(0, y - \text{Std}(z_j))$ penalizza le varianze delle singole dimensioni che scendono sotto una soglia y , prevenendo il collasso dimensionale:

$$\text{Std}(z_j) = \sqrt{(\text{Var}(z_j) + \epsilon)} = \sqrt{((1/N) \sum_i (z_{\{i,j\}} - \bar{z}_j)^2 + \epsilon)}$$

Il termine di covarianza $\mathcal{L}_{\text{cov}} = (1/(d(d-1))) \sum_{\{i \neq j\}} [\text{Cov}(z)_{\{i,j\}}]^2$ penalizza le correlazioni tra dimensioni diverse, incoraggiando la decorrelazione delle componenti della rappresentazione — un obiettivo analogo all'analisi delle

componenti principali (PCA) ma ottimizzato end-to-end con il resto della perdita.

La scelta tra queste strategie di prevenzione del collasso non è puramente ingegneristica: riflette ipotesi implicite sulla struttura dello spazio latente ottimale. I metodi basati su varianza-covarianza (VICReg, BarlowTwins) assumono implicitamente che le dimensioni della rappresentazione debbano essere scorrelate e distribuite uniformemente — un'assunzione analoga all'ipotesi di indipendenza statistica dell'ICA (Independent Component Analysis). I metodi basati su EMA (BYOL, I-JEPA) non assumono una struttura particolare sulla distribuzione delle rappresentazioni, lasciando che questa emerga dalla struttura del compito di predizione. La questione di quale assunzione sia più appropriata per l'apprendimento di rappresentazioni causali del mondo fisico rimane aperta e costituisce uno dei filoni di ricerca più attivi nel campo.

6. Genealogia del concetto di Modello di Mondo

6.1 Kenneth Craik e la nascita dell'ipotesi dei modelli interni

La storia intellettuale del concetto di modello di mondo non può essere narrata senza iniziare da Kenneth James William Craik, lo psicologo scozzese che nel 1943, a soli ventinove anni e con una monografia di appena centoventi pagine, pose le fondamenta di un'idea destinata a ridisegnare il paesaggio della scienza cognitiva, delle neuroscienze computazionali e dell'intelligenza artificiale per i decenni successivi. "The Nature of Explanation" (Cambridge University Press, 1943) è un testo anomalo nella letteratura scientifica: non propone esperimenti, non avanza teorie matematizzate, non discute in modo sistematico dati empirici. È invece un saggio filosofico sulla natura della spiegazione scientifica che, quasi en passant, formula un'ipotesi straordinariamente audace: che il sistema nervoso centrale operi attraverso la costruzione di modelli interni della realtà.

La tesi di Craik, nei suoi termini essenziali, è questa. Il cervello non risponde agli stimoli del mondo in modo diretto, come un sistema riflesso; esso porta con sé una rappresentazione in piccola scala — un modello — della realtà esterna e delle proprie azioni possibili. Questo modello permette di ragionare su eventi, di anticipare le conseguenze delle azioni prima di eseguirle, di scegliere tra alternative senza doverle realizzare fisicamente. La capacità di previsione — di simulare internamente il futuro — è, secondo Craik, il tratto distintivo dell'intelligenza come fenomeno biologico, ciò che distingue la cognizione dal semplice riflesso.

L'originalità di questa formulazione deve essere apprezzata nel suo contesto storico. Nel 1943 il campo era dominato dal comportamentismo di Watson e Skinner, che si rifiutava programmaticamente di postulare stati interni non osservabili e riduceva la cognizione a relazioni di stimolo-risposta. Parlare di "modelli interni" era, in quel clima intellettuale, una provocazione epistemologica. Craik era consapevole di questa

tensione e vi rispondeva argomentativamente: se il sistema nervoso deve rispondere in modo adattivo a eventi futuri che non si sono ancora verificati, deve necessariamente possedere qualcosa che svolga la funzione di una rappresentazione di quegli eventi. Il modello interno non è un'ipotesi metafisica ma un'inferenza funzionale.

La formulazione di Craik era verbale, non matematica, e rimase tale fino alla sua morte prematura nel 1945, investito da un'automobile mentre si recava in bicicletta a un picnic il giorno del VE Day. Non poté mai sapere che le sue intuizioni avrebbero attraversato il successivo mezzo secolo come corrente sotterranea, per riemergere con forza negli anni Ottanta nella psicologia cognitiva di Philip Johnson-Laird, negli anni Novanta nel predictive coding computazionale di Rao e Ballard, e negli anni Duemila nel Free Energy Principle di Karl Friston.

6.2 La psicologia cognitiva e le mappe cognitive di Tolman

Mentre Craik formulava la sua ipotesi sul versante filosofico, un'intuizione convergente emergeva dal lato sperimentale della psicologia americana. Edward Chace Tolman, in una serie di esperimenti di labirinto condotti negli anni Trenta e Quaranta con ratti di laboratorio, aveva accumulato evidenze che ponevano in seria difficoltà l'interpretazione comportamentista dell'apprendimento come semplice concatenazione di associazioni stimolo-risposta.

Il risultato più celebre è il paradigma del latent learning. Ratti che avevano esplorato un labirinto senza ricompensa — che non avevano dunque ricevuto rinforzo per nessuna sequenza di mosse — mostravano, quando veniva introdotta una ricompensa nella meta, una capacità di raggiungere l'obiettivo significativamente superiore rispetto ai controlli che non avevano esplorato. L'esplorazione aveva prodotto qualcosa che il comportamentismo non poteva spiegare: una conoscenza della struttura del labirinto non associata a nessuna risposta motoria specifica, una mappa cognitiva che poteva essere sfruttata flessibilmente al momento opportuno.

Tolman denominò questo fenomeno apprendimento latente e coniò il termine cognitive map — mappa cognitiva — per descrivere la struttura interna che il ratto aveva costruito attraverso l'esplorazione. La mappa cognitiva di Tolman non è una sequenza di istruzioni motorie né una lista di associazioni stimolo-risposta: è una rappresentazione della topologia dello spazio che può essere interrogata per generare percorsi verso qualsiasi destinazione, compresi percorsi mai percorsi prima. Nel 1948, nel suo discorso presidenziale all'American Psychological Association, Tolman scrisse che le mappe cognitive sono rappresentazioni dello spazio fisico che l'organismo porta con sé e che usa per navigare.

La distinzione tra la conoscenza della struttura dell'ambiente (il modello) e la conoscenza di quale azione eseguire in quale stato (la policy) è una distinzione che il reinforcement learning moderno ha formalizzato matematicamente attraverso l'opposizione tra metodi model-based e metodi model-free. Tolman aveva intuito questa distinzione decenni prima che il formalismo teorico fosse disponibile per articolarla con precisione.

Philip Johnson-Laird, nel suo "Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness" (1983), ha riformulato l'intuizione di Craik in termini psicologici più precisi. Johnson-Laird argomenta che il ragionamento umano non avviene attraverso la manipolazione di regole formali logiche — come sosteneva la logica degli psicologi di tradizione chomskiana — ma attraverso la costruzione e la manipolazione di modelli mentali: rappresentazioni analogiche delle situazioni che preservano le relazioni strutturali tra gli elementi. Per ragionare sulla validità di un sillogismo, le persone costruiscono un modello mentale della situazione descritta dalle premesse e verificano se la conclusione è vera in tutti i modelli compatibili con quelle premesse.

L'aspetto che connette più direttamente Johnson-Laird alla proposta di LeCun è l'enfasi sulla natura continua e analogica, anziché discreta e simbolica, dei modelli mentali. Le rappresentazioni interne non sono stringhe di simboli manipolabili attraverso regole formali ma strutture che preservano le relazioni spaziali, causali e temporali tra gli elementi — strutture, cioè, che hanno una geometria. Questa concezione si allinea con l'idea lecuniana di predizione in spazio latente continuo, in netto contrasto con la manipolazione di token discreti che caratterizza i sistemi linguistici autoregressivi.

6.3 Il Modello di Mondo nel reinforcement learning: formalismo e architetture

La formalizzazione matematica del concetto di modello di mondo all'interno della teoria del reinforcement learning richiede di richiamare il formalismo del processo decisionale markoviano (Markov Decision Process, MDP), che costituisce il quadro teorico entro cui si articola la maggior parte della ricerca attuale sull'apprendimento per rinforzo.

Un MDP è una tupla $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ dove $\mathcal{S}$ è lo spazio degli stati, $\mathcal{A}$ è lo spazio delle azioni, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0,1]$ è la funzione di transizione di stato (con $P(s' | s, a)$ la probabilità di transitare nello stato s' dati lo stato corrente s e l'azione a), $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ è la funzione di ricompensa istantanea, e $\gamma \in [0,1)$ è il fattore di sconto temporale. La proprietà di Markov — che il futuro sia condizionalmente indipendente dal passato dato lo stato corrente — è l'assunzione fondamentale che rende il problema computazionalmente trattabile.

L'obiettivo di un agente di reinforcement learning è trovare una policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ (o una distribuzione di probabilità sulle azioni $\pi(a|s)$) che massimizzi il ritorno atteso scontato:

$$G_t = \sum_{k=0}^{\infty} \gamma^k \cdot R(s_{t+k}, a_{t+k})$$

Il valore di uno stato s sotto la policy π è definito come il ritorno atteso a partire da s seguendo π :

$$V^{\pi}(s) = \mathbb{E}_{\pi} [G_t | s_t = s] = \mathbb{E}_{\pi} [\sum_{k=0}^{\infty} \gamma^k \cdot R(s_{t+k}, a_{t+k}) | s_t = s]$$

La funzione di valore ottima $V^*(s) = \max_{\pi} V^{\pi}(s)$ soddisfa le equazioni di Bellman per il valore ottimale:

$$V^*(s) = \max_{a \in \mathcal{A}} [R(s,a) + \gamma \cdot \sum_{s'} P(s' | s,a) \cdot V^*(s')]]$$

Analogamente, la funzione Q-value ottimale $Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$, che misura il valore atteso di eseguire l'azione a nello stato s e poi seguire la policy ottimale, soddisfa:

$$Q^*(s,a) = R(s,a) + \gamma \cdot \sum_{s'} P(s' | s,a) \cdot \max_{a'} Q^*(s',a')$$

La distinzione cruciale tra metodi model-free e model-based riguarda l'utilizzo esplicito della funzione di transizione $P(s' | s,a)$ e della funzione di ricompensa $R(s,a)$.

I metodi model-free — Q-learning (Watkins e Dayan, 1992), REINFORCE (Williams, 1992), Proximal Policy Optimization (PPO, Schulman et al., 2017), Soft Actor-Critic (SAC, Haarnoja et al., 2018) — stimano direttamente la funzione di valore o la policy attraverso l'esperienza, senza mai costruire esplicitamente un modello dell'ambiente. Il Q-learning, nella sua forma tabulare, aggiorna la stima $Q(s,a)$ attraverso la regola:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot [r + \gamma \cdot \max_{a'} Q(s',a') - Q(s,a)]$$

dove α è il tasso di apprendimento e il termine tra parentesi quadre è l'errore di predizione temporale (temporal difference error, TD error). Questa regola non richiede la conoscenza di $P(s' | s,a)$ o $R(s,a)$: impara direttamente dalle transizioni osservate (s,a,r,s') . Il vantaggio di questa strategia è la semplicità e la generale applicabilità; il costo principale è la scarsa efficienza campionaria: un agente puramente model-free deve osservare molte interazioni con l'ambiente per stimare accuratamente la funzione di valore, un problema che diventa critico in domini in cui le interazioni sono costose o pericolose.

I metodi model-based — Dyna (Sutton, 1990), PILCO (Deisenroth e Rasmussen, 2011), MuZero (Schrittwieser et al., 2020) — imparano esplicitamente un modello dell'ambiente — una funzione di transizione $P(s' | s,a)$ e/o una funzione di ricompensa $R(s,a)$ — e lo usano per migliorare la policy attraverso pianificazione o simulazione.

L'architettura Dyna di Sutton (1990) è il prototipo concettuale di questa famiglia. Il suo ciclo di apprendimento alterna due tipi di aggiornamenti: aggiornamenti diretti attraverso l'esperienza reale con l'ambiente (identici al Q-learning puro), e aggiornamenti di pianificazione attraverso interazioni simulate con il modello appreso P . Formalmente, dato un modello P appreso dalle transizioni reali, l'agente esegue n passi di pianificazione per ogni interazione reale campionando stati e azioni dal passato e generando transizioni simulate:

per $i = 1, \dots, n$: campiona s uniformemente dagli stati visitati campiona a uniformemente dalle azioni eseguite in s ottieni $s', \hat{r} = P(s,a), R(s,a)$ $Q(s,a) \leftarrow Q(s,a) + \alpha \cdot [\hat{r} + \gamma \cdot \max_{a'} Q(s',a') - Q(s,a)]$

L'effetto di questo meccanismo è una drastica riduzione del numero di interazioni reali necessarie per convergere a una policy competente: ogni transizione reale genera n transizioni simulate che allenano ulteriormente la funzione di valore. Il rapporto di efficienza campionaria tra Dyna e Q-learning puro cresce linearmente con n , ma ovviamente a condizione che il modello appreso P sia accurato — errori sistematici nel modello si propagano agli aggiornamenti di pianificazione e possono portare a policy subottimali.

PILCO (Probabilistic Inference for Learning COntrol, Deisenroth e Rasmussen, 2011) affronta il problema dell'accuratezza del modello in modo più sofisticato, utilizzando processi gaussiani (Gaussian Processes, GP) come modello dell'ambiente. Un GP è un processo stocastico la cui distribuzione marginale su qualunque insieme finito di punti è gaussiana multivariata, e costituisce un prior non parametrico sulle funzioni. Il modello di transizione del GP è:

$$s_{t+1} = f(s_t, a_t) + \epsilon_t, \epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon)$$

con $f \sim \text{GP}(\mu, k)$ dove $\mu : \mathbb{R}^{(n+m)} \rightarrow \mathbb{R}^n$ è la funzione di media e $k : \mathbb{R}^{(n+m)} \times \mathbb{R}^{(n+m)} \rightarrow \mathbb{R}$ è la funzione kernel (tipicamente il kernel RBF: $k(x, x') = \sigma^2 \exp(-\|x - x'\|^2 / (2\ell^2))$). La previsione del GP in un punto x^* non osservato è:

$$p(f(x^*) \mid \mathcal{D}) = \mathcal{N}(\mu^*(x^*), \sigma^{2*}(x^*))$$

con $\mu^*(x^*) = \mu(x^*) + k(x^*, X) [K(X, X) + \Sigma_\epsilon]^{-1} (y - \mu(X))$ e $\sigma^{2*}(x^*) = k(x^*, x^*) - k(x^*, X) [K(X, X) + \Sigma_\epsilon]^{-1} k(X, x^*)$, dove X sono i punti osservati, y i corrispondenti valori target, e $K(X, X)$ è la matrice di Gram. Il vantaggio del GP è che fornisce stime di incertezza calibrate: il modello sa quanto non sa, e questa incertezza può essere propagata attraverso la pianificazione per produrre policy conservative nelle regioni dello spazio degli stati poco esplorate.

MuZero (Schrittwieser et al., 2020) di DeepMind rappresenta l'approccio model-based più avanzato nella tradizione del reinforcement learning profondo. A differenza di PILCO, non apprende un modello del mondo nello spazio degli stati osservati, ma un modello di transizione latente che opera direttamente nello spazio delle rappresentazioni apprese. L'architettura si compone di tre reti neurali: una funzione di rappresentazione $h_\theta : \mathcal{S}_{\text{obs}} \rightarrow \mathbb{R}^d$ che mappa le osservazioni grezze in stati latenti, una funzione di dinamica $g_\theta : \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R}^d \times \mathbb{R}$ che predice il prossimo stato latente e la ricompensa, e una funzione di predizione $f_\theta : \mathbb{R}^d \rightarrow \Delta(\mathcal{A}) \times \mathbb{R}$ che mappa gli stati latenti in distribuzioni di policy e stime di valore. L'addestramento avviene attraverso backpropagation attraverso le predizioni del modello unroll per K passi:

$$\mathcal{L}(\theta) = \sum_{k=0}^{K-1} [\mathcal{L}_r(r^k, \hat{r}^k) + \mathcal{L}_v(z^k, \hat{v}^k) + \mathcal{L}_p(\pi^k, \hat{p}^k)]$$

dove r^k è la ricompensa reale al passo k , $\hat{r}^k$ è la ricompensa predetta dal modello, z^k è il target di valore stimato da MCTS (Monte Carlo Tree Search), $\hat{v}^k$ è il valore predetto, π^k è la policy MCTS, e $\hat{p}^k$ è la policy predetta. Le perdite $\mathcal{L}_r$, $\mathcal{L}_v$, $\mathcal{L}_p$ sono rispettivamente cross-entropy e mean squared error.

MuZero è rilevante per la proposta di LeCun in due modi. Primo, dimostra che i modelli di mondo in spazio latente possono essere appresi con successo e usati per pianificazione efficace in domini complessi (Go, Chess, Shogi, Atari). Secondo, mostra che il modello non ha bisogno di essere un modello completo del mondo fisico — non deve predire i pixel del prossimo frame — ma solo le quantità necessarie per la pianificazione: la ricompensa e il valore atteso. Questa "sufficienza funzionale" del modello di transizione latente è precisamente ciò che LeCun sottolinea con il principio della predizione in spazio latente.

La differenza principale tra MuZero e la proposta lecuniana è l'obiettivo di apprendimento: MuZero impara il modello attraverso il rinforzo (ottimizzando la policy per massimizzare la ricompensa cumulata), mentre LeCun propone di impararlo attraverso l'osservazione autosupervisionata del mondo fisico, senza rinforzo esplicito. Questa distinzione riflette una differenza filosofica profonda: il modello di LeCun deve essere una rappresentazione della struttura del mondo fisico acquisita per osservazione, non un modello ottimizzato per supportare una particolare funzione di ricompensa.

6.4 PILCO e i Gaussian World Models: propagazione analitica dell'incertezza

Vale la pena approfondire il contributo di PILCO all'idea di modello di mondo, perché introduce una questione formalmente rilevante per la proposta lecuniana: come propagare l'incertezza del modello attraverso la pianificazione in modo da ottenere stime di policy che riflettano onestamente ciò che il sistema non sa.

In PILCO, la pianificazione avviene attraverso la propagazione delle distribuzioni di stato attraverso il modello GP. Dato lo stato iniziale $s_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$ e la policy π (tipicamente un controllo lineare gaussiano $\pi(s) = W \cdot s + b + \epsilon$ con $\epsilon \sim \mathcal{N}(0, \sigma^2 \pi)$), il prossimo stato s_1 è:

$$p(s_1 | s_0, a_0) = \int p(s_1 | f) p(f | s_0, a_0) df$$

dove la marginalizzazione sulle previsioni del GP produce una distribuzione che non è in generale gaussiana — la composizione di una gaussiana sull'input con un GP produce output non gaussiani. PILCO approssima questa distribuzione con una gaussiana $p(s_1) \approx \mathcal{N}(\mu_1, \Sigma_1)$ attraverso moment matching: si calcolano il valore atteso e la covarianza di s_1 analiticamente sfruttando le proprietà del kernel RBF, e si approssima la distribuzione con la gaussiana corrispondente.

Questa approssimazione viene iterata per T passi di pianificazione, producendo una sequenza di gaussiane approssimate $\{\mathcal{N}(\mu_t, \Sigma_t)\}_{t=0}^T$. La policy ottimale viene trovata ottimizzando il costo atteso:

$$J(\pi) = \sum_{t=0}^T \mathbb{E}_{s_t \sim \mathcal{N}(\mu_t, \Sigma_t)} [c(s_t)]$$

attraverso gradient descent rispetto ai parametri della policy, dove il gradiente è calcolato analiticamente attraverso le espressioni closed-form del GP.

La potenza di PILCO è la sua eccezionale efficienza campionaria: in domini di controllo di media complessità, può imparare policy competenti da una dozzina di interazioni reali — un record che i metodi model-free non avvicinano. Il suo limite principale è la scalabilità: la complessità computazionale dell'inferenza in un GP cresce cubicamente con il numero di punti di addestramento (a causa dell'inversione della matrice di Gram $[K(X,X) + \Sigma_\epsilon]^{-1}$, operazione $O(N^3)$), rendendo PILCO impraticabile per problemi con molti dati e alta dimensionalità dello spazio degli stati.

LeCun cita PILCO e i suoi successori come prove di concetto dell'idea che modelli di mondo accurati possano drasticamente migliorare l'efficienza campionaria dell'apprendimento — ma la sua proposta aspira a scalare a domini di alta dimensionalità attraverso reti neurali profonde, non a sfruttare la potenza delle stime di incertezza bayesiane che i GP offrono. Il trade-off tra qualità delle stime di incertezza e scalabilità è una delle tensioni aperte nel campo.

6.5 Ha e Schmidhuber: World Models (2018) — architettura e risultati

Il lavoro di David Ha e Jürgen Schmidhuber "World Models" (2018, arXiv:1803.10122) rappresenta l'antecedente diretto più rilevante della proposta di LeCun, non tanto per la convergenza nelle idee architetture — le differenze sono significative — quanto per l'influenza nella definizione dell'agenda di ricerca e nel dimostrare empiricamente la praticabilità dell'idea di apprendimento di un modello di mondo in spazio latente per la pianificazione.

L'architettura di Ha e Schmidhuber separa nettamente la percezione, il modello del mondo e il controller in tre componenti modulari.

Il componente di percezione è un Variational Autoencoder (VAE, Kingma e Welling, 2013) convoluzionale che mappa ogni frame video $x_t \in \mathbb{R}^{(H \times W \times 3)}$ in una distribuzione gaussiana nello spazio latente:

$$q_\varphi(z_t | x_t) = \mathcal{N}(\mu_\varphi(x_t), \text{diag}(\sigma_\varphi^2(x_t)))$$

dove μ_φ e σ_φ^2 sono le uscite dell'encoder. Il campione latente $z_t \sim q_\varphi(z_t | x_t)$ viene ottenuto attraverso il reparametrization trick: $z_t = \mu_\varphi(x_t) + \sigma_\varphi(x_t) \odot \epsilon_t$ con $\epsilon_t \sim \mathcal{N}(0, I)$. Il VAE è addestrato minimizzando l'Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{VAE}}(\varphi, \psi) = \mathbb{E}_{\{z_t \sim q_\varphi\}} [\log p_\psi(x_t | z_t)] - \text{KL}[q_\varphi(z_t | x_t) \parallel p(z_t)]$$

dove $p_\psi(x_t | z_t)$ è il decoder che ricostruisce il frame dal vettore latente e $p(z_t) = \mathcal{N}(0, I)$ è il prior gaussiano standard. Il termine KL regolarizza la distribuzione variazionale verso il prior, assicurando che lo spazio latente abbia una struttura regolare e che campioni di questo prior possano essere decodificati in immagini plausibili.

Il modello del mondo è una Mixture Density Network (MDN) costruita sopra una Recurrent Neural Network (RNN), specificatamente una Long Short-Term Memory (LSTM, Hochreiter e Schmidhuber, 1997). La MDN-RNN mantiene uno stato nascosto h_t che viene aggiornato a ogni passo:

$$h_t = \text{LSTM}(z_t, a_t, h_{t-1})$$

e usa questo stato per predire la distribuzione condizionale sul prossimo stato latente z_{t+1} come mixture di gaussiane:

$$p(z_{t+1} \mid z_t, a_t, h_t) = \sum_{k=1}^K \pi_k(h_t) \cdot \mathcal{N}(z_{t+1}; \mu_k(h_t), \sigma_k^2(h_t))$$

dove π_k , μ_k , σ_k sono le probabilità di mixing, le medie e le deviazioni standard delle K componenti gaussiane, tutti output della MDN-RNN. La scelta di una mixture di gaussiane come distribuzione di uscita permette di modellare distribuzioni multimodali — situazioni in cui il futuro è genuinamente ambiguo tra più evoluzioni distinte — che una singola gaussiana non potrebbe catturare.

Il controller è volutamente minimalista: una semplice rete lineare che mappa lo stato latente z_t e lo stato nascosto h_t nell'azione:

$$a_t = W_c \cdot [z_t; h_t] + b_c$$

con $W_c \in \mathbb{R}^{(d_a \times (d_z + d_h))}$ e $b_c \in \mathbb{R}^{d_a}$. I parametri del controller vengono ottimizzati attraverso evoluzione stocastica (CMA-ES, Covariance Matrix Adaptation - Evolution Strategy, Hansen, 2006) anziché gradient descent — una scelta che permette di ottimizzare anche funzioni di fitness non differenziabili e di evitare i problemi di ottimizzazione del gradiente attraverso sequenze temporali lunghe.

Il contributo più significativo del paper di Ha e Schmidhuber è la dimostrazione che il controller può essere addestrato interamente "nella testa" del modello del mondo: invece di interagire con l'ambiente simulato vero (un game engine), l'agente genera traiettorie nella simulazione prodotta dalla propria MDN-RNN e allena il controller su queste traiettorie simulate. La policy risultante, trasferita poi all'ambiente reale, mostra una qualità competitiva con policy addestrate direttamente sull'ambiente reale — una validazione empirica notevole dell'utilità della pianificazione in spazio latente.

La differenza fondamentale rispetto alla proposta di LeCun è il paradigma di apprendimento: Ha e Schmidhuber operano nel quadro del reinforcement learning model-based, con una funzione di ricompensa esplicita che guida l'ottimizzazione del controller. LeCun propone invece che il modello di mondo venga appreso attraverso osservazione autosupervisionata passiva, senza rinforzo, e che la pianificazione avvenga attraverso gradient descent sulla funzione di costo differenziabile — non attraverso algoritmi evolutivi. Questa differenza non è tecnica ma concettuale: riflette una visione radicalmente diversa di come l'intelligenza debba acquisire la propria conoscenza del mondo.

6.6 MuZero e il modello di mondo per il gioco: pianificazione in spazio latente con MCTS

MuZero (Schrittwieser et al., 2020) merita una discussione separata perché realizza con straordinario successo empirico l'idea di pianificazione attraverso un modello di transizione in spazio latente, dimostrando che questo approccio supera anche i metodi model-free di ultima generazione su compiti altamente complessi.

Il cuore di MuZero è l'utilizzo del modello di transizione latente g_θ all'interno di una ricerca ad albero Monte Carlo (MCTS). Per ogni stato corrente, l'agente esegue una simulazione dell'albero dei giochi possibili proiettando lo stato corrente s nello spazio latente attraverso h_θ , e poi simulando K passi di gioco attraverso g_θ e f_θ . In ogni nodo dell'albero, il valore atteso e la distribuzione di policy prodotti da f_θ guidano la ricerca verso rami promettenti. Dopo N simulazioni, l'agente esegue l'azione suggerita dalle statistiche dell'albero.

Formalmente, l'algoritmo MCTS in MuZero seleziona le azioni in ogni nodo dell'albero usando il criterio UCB (Upper Confidence Bound) modificato per MCTS:

$$a_t = \operatorname{argmax}_a \{ [Q(s,a) + U(s,a)]$$

con $U(s,a) = C(s) \cdot \hat{p}(a|s) \cdot \sqrt{N(s)} / (1 + N(s,a))$, dove $Q(s,a)$ è la stima di valore dell'azione a nello stato s (media degli ritorni osservati in tutte le simulazioni che hanno attraversato il nodo (s,a)), $N(s)$ e $N(s,a)$ sono rispettivamente il numero totale di visite al nodo s e al nodo (s,a) , $\hat{p}(a|s)$ è la probabilità dell'azione a predetta dalla rete di policy f_θ , e $C(s)$ è un fattore di esplorazione.

Il notevole risultato di MuZero — che eguaglia o supera AlphaZero (che ha accesso alle regole del gioco) su Go, Chess, Shogi e supera tutti i sistemi precedenti su Atari — dimostra che modelli di mondo in spazio latente appresi da esperienza possono essere sufficientemente accurati per pianificazione efficace anche in ambienti di alta complessità. Tuttavia, MuZero opera in un dominio in cui le interazioni sono esatte (nessun rumore ambientale, stato completamente osservabile nei giochi da tavolo) e la funzione di ricompensa è precisa e definita esternamente. La generalizzazione a domini fisici reali — dove le osservazioni sono rumorose, lo stato è parzialmente osservabile, e la funzione di costo deve essere appresa — è la sfida che LeCun affronta con la proposta JEPA.

6.7 L'interpretazione lecuniana: tre trasformazioni concettuali

LeCun riprende il concetto di Modello di Mondo da questa genealogia ricca e la trasforma in almeno tre modi che ne alterano il carattere profondamente.

La prima trasformazione riguarda il paradigma di apprendimento. Tutti i lavori precedentemente citati — da Dyna a PILCO, da Ha-Schmidhuber a MuZero — apprendono il modello di mondo attraverso interazione diretta con l'ambiente, con o senza ricompensa esplicita. LeCun propone invece un modello di mondo appreso attraverso osservazione autosupervisionata passiva di grandi quantità di dati del

mondo fisico — video, immagini, segnali sensoriali — senza nessuna interazione, nessun rinforzo, nessuna segnalazione di errore dall'ambiente. Il modello di mondo non è costruito per supportare un compito specifico: è una struttura di conoscenza generale del mondo fisico che può poi essere specializzata per diversi compiti attraverso fine-tuning o prompting.

Questa scelta è motivata dal paragone con lo sviluppo cognitivo biologico. I bambini costruiscono la loro comprensione della fisica degli oggetti — della permanenza, della solidità, della gravità, della causalità meccanica — principalmente attraverso osservazione passiva e interazione spontanea, non attraverso un programma strutturato di ricompense. Il modello di mondo di LeCun aspira a realizzare lo stesso tipo di acquisizione autosupervisionata.

La seconda trasformazione riguarda lo spazio di predizione. Mentre i lavori precedenti oscillano tra la predizione nello spazio osservativo (Ha e Schmidhuber prevedono la distribuzione su z ma l'addestrando rimane ancorato alla ricostruzione dei frame attraverso il VAE) e la predizione in spazio latente puro (MuZero), LeCun pone la predizione in spazio latente come principio architetturale assoluto: non si deve mai cercare di predire lo spazio delle osservazioni grezze, perché questo spazio contiene informazioni irriducibilmente imprevedibili che distorcono l'apprendimento verso la modellazione del rumore anziché della struttura causale.

La terza trasformazione riguarda l'integrazione in un'architettura modulare completa. LeCun non propone solo un modello di mondo isolato ma un'architettura che integra il modello di mondo con un modulo percettivo, una memoria a lungo termine, un modulo di costo esplicito, un pianificatore gerarchico e un modulo attoriale. Questa architettura aspira a realizzare il ciclo completo della cognizione — dalla percezione alla pianificazione all'azione — in un sistema unificato ma modulare che riflette, nella sua struttura funzionale, l'organizzazione della cognizione biologica.

7. Fondamenti Cognitivi e Neuroscientifici: il Free Energy Principle di Friston

7.1 Le radici storiche del predictive coding: da Helmholtz a Rao e Ballard

La teoria del predictive coding — la percezione come processo di inferenza attiva piuttosto che di estrazione passiva di feature — ha radici che risalgono all'Ottocento, alla figura di Hermann von Helmholtz, il fisiologo e fisico tedesco che ha dominato le scienze della percezione nella seconda metà del XIX secolo. Helmholtz aveva proposto che la percezione non fosse una registrazione diretta dello stimolo sensoriale ma un processo di inferenza inconscia: il sistema nervoso centrale costruisce la migliore interpretazione delle cause degli stimoli sensoriali basandosi sulla loro

struttura e sulle conoscenze pregresse, in un processo che Helmholtz paragonò all'inferenza bayesiana inconscia.

L'idea helmholtziana rimase per un secolo un'intuizione filosofica senza formalizzazione matematica precisa. La formalizzazione computazionale arrivò nel 1999 con il lavoro seminale di Rajesh Rao e Dana Ballard, "Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-Field Effects", pubblicato su Nature Neuroscience. Rao e Ballard proposero un modello computazionale della corteccia visiva gerarchica in cui ogni livello della gerarchia mantiene un modello generativo delle attivazioni del livello inferiore, e comunica con il livello inferiore attraverso due tipi di segnali: predizioni discendenti (top-down) che rappresentano le aspettative del livello superiore sulle attivazioni del livello inferiore, ed errori di predizione ascendenti (bottom-up) che rappresentano la discrepanza tra le predizioni e le attivazioni effettive.

La formulazione matematica di Rao e Ballard per un modello a due livelli è la seguente. Siano r^1 le attivazioni al livello 1 (il livello inferiore, più vicino alle entrate sensoriali) e r^2 le attivazioni al livello 2 (il livello superiore). Il modello generativo ipotizza che r^1 sia generato da r^2 attraverso una funzione di generazione f :

$$r^1 = f(r^2) + \epsilon^1, \epsilon^1 \sim \mathcal{N}(0, \Sigma^1)$$

dove ϵ^1 è rumore gaussiano con covarianza Σ^1 . L'energia di inferenza — la quantità che il sistema tende a minimizzare — è:

$$E = \frac{1}{2} (r^1 - f(r^2))^T (\Sigma^1)^{-1} (r^1 - f(r^2)) + \frac{1}{2} (r^2 - r^2_{\text{prior}})^T (\Sigma^2)^{-1} (r^2 - r^2_{\text{prior}})$$

dove r^2_{prior} è il prior sulle attivazioni del livello 2. La minimizzazione di E rispetto a r^2 (inferenza) produce le equazioni di aggiornamento:

$$\partial r^2 / \partial t = -\partial E / \partial r^2 = (\partial f / \partial r^2)^T (\Sigma^1)^{-1} (r^1 - f(r^2)) - (\Sigma^2)^{-1} (r^2 - r^2_{\text{prior}})$$

Il termine $(r^1 - f(r^2))$ è l'errore di predizione al livello 1 — la discrepanza tra le attivazioni osservate e quelle predette dal modello generativo di livello 2. Il termine $(\partial f / \partial r^2)^T (\Sigma^1)^{-1}$ è il jacobiano della funzione generativa pesato con la precisione, che converte gli errori di predizione al livello inferiore in aggiornamenti per le credenze al livello superiore. Questo schema computazionale produce naturalmente i due tipi di connessioni osservati nella corteccia visiva: le connessioni ricorrenti top-down che trasmettono le predizioni $(f(r^2))$ e le connessioni feed-forward bottom-up che trasmettono gli errori di predizione $(r^1 - f(r^2))$.

Il modello di Rao e Ballard spiegava elegantemente alcuni effetti fisiologici precedentemente misteriosi, in particolare la soppressione di risposta alle stimolazioni attese (le cosiddette extra-classical receptive field effects): le aree corticali primarie rispondono con minore intensità agli stimoli che il modello generativo si aspetta, e con maggiore intensità agli stimoli anomali — esattamente ciò che predice un sistema che trasmette solo gli errori di predizione e non gli interi pattern sensoriali.

7.2 Il Free Energy Principle: derivazione formale rigorosa

Karl Friston ha sviluppato il Free Energy Principle (FEP) nel corso di due decenni di lavoro, a partire da una serie di articoli pubblicati su *Journal of Physiology* e *NeuroImage* tra il 2003 e il 2010, convergendo nella formulazione più ambiziosa e generale in "The free-energy principle: a unified brain theory?" (*Nature Reviews Neuroscience*, 2010). L'ambizione della proposta è senza precedenti nella neuroscienze teoriche: formulare un principio unico che governi tutti i sistemi auto-organizzati — dal singolo neurone all'intero cervello, dall'agente biologico alla colonia di insetti.

Il FEP si fonda sull'osservazione che i sistemi viventi si mantengono in stati lontani dall'equilibrio termodinamico occupando una nicchia ecologica circoscritta: la distribuzione dei loro stati sensoriali è limitata, non uniforme, e questo fatto distingue un essere vivente dalla materia inanimata. Formalmente, un essere vivente è caratterizzato da una distribuzione $q(s)$ sui propri stati sensoriali che è concentrata su un sottoinsieme limitato dello spazio degli stati — gli stati compatibili con la sopravvivenza. La sorpresa di uno stato sensoriale s è:

$$I(s) = -\log p(s)$$

dove $p(s)$ è la probabilità dell'osservazione s sotto un modello generativo implicito del mondo. La sopravvivenza richiede che la sorpresa media — l'entropia della distribuzione sensoriale — sia bassa:

$$H(p) = \mathbb{E}_{\{s \sim p(s)\}} [I(s)] = -\mathbb{E}_{\{s \sim p(s)\}} [\log p(s)]$$

Il problema computazionale fondamentale è che il cervello non ha accesso diretto a $p(s)$ — la vera distribuzione delle cause sensoriali nel mondo — e non può quindi calcolare $I(s)$ direttamente. Il FEP risolve questo problema attraverso l'introduzione dell'energia libera variazionale F , una quantità accessibile al sistema che fornisce un limite superiore alla sorpresa:

$$F = \mathbb{E}_{\{q(z)\}} [\log q(z) - \log p(s, z)] = \text{KL}[q(z) \parallel p(z | s)] - \log p(s)$$

dove z sono le cause latenti degli stati sensoriali s , $q(z)$ è la distribuzione variazionale approssimata (le "credenze" del cervello sulle cause latenti), $p(s, z)$ è il modello generativo del cervello (la distribuzione congiunta di cause e effetti secondo il modello interno del mondo), e $p(z | s)$ è la distribuzione di posterior vera sulle cause dati gli stati sensoriali.

La fattorizzazione fondamentale dell'energia libera può essere scritta in due modi complementari, ognuno dei quali illumina un aspetto diverso del principio:

Fattorizzazione 1 (inferenza-predizione):

$$F = \text{KL}[q(z) \parallel p(z | s)] - \log p(s) \geq -\log p(s) = I(s)$$

Questa fattorizzazione mostra che F è sempre maggiore o uguale alla sorpresa $I(s)$, con uguaglianza quando $q(z) = p(z|s)$ — ovvero quando le credenze del cervello corrispondono esattamente alla distribuzione di posterior vera. La minimizzazione di F rispetto a $q(z)$ realizza pertanto l'inferenza bayesiana ottimale: rende le credenze del cervello sulle cause latenti il più vicino possibile alla distribuzione di posterior vera, minimizzando così il gap tra F e $I(s)$.

Fattorizzazione 2 (energia-entropia):

$$F = \mathbb{E}_{q(z)} [-\log p(s,z)] - H(q(z)) = U(q) - H(q)$$

dove $U(q) = \mathbb{E}_{q(z)} [-\log p(s,z)]$ è l'energia attesa (equivalente all'energia potenziale in fisica statistica) e $H(q) = -\mathbb{E}_{q(z)} [\log q(z)]$ è l'entropia della distribuzione variazionale. Questa fattorizzazione mostra che minimizzare F equivale a minimizzare l'energia attesa (rendere plausibili le stime delle cause latenti sotto il modello generativo) mantenendo alta l'entropia della distribuzione variazionale (preservando l'incertezza dove essa è appropriata).

Fattorizzazione 3 (accuratezza-complessità):

$$F = \mathbb{E}_{q(z)} [-\log p(s|z)] + \text{KL}[q(z) \parallel p(z)]$$

dove il primo termine è l'energia di ricostruzione — quanto il modello generativo $p(s|z)$ spiega l'osservazione s sotto le credenze correnti $q(z)$ — e il secondo è la complessità variazionale — la distanza tra le credenze posteriori $q(z)$ e il prior $p(z)$. Questa fattorizzazione è quella che appare nel training degli autoencoder variazionali (VAE) come ELBO negativo: la perdita di ricostruzione e il termine KL di regolarizzazione. La minimizzazione di F come ELBO è dunque formalmente equivalente alla minimizzazione dell'energia libera variazionale nel senso del FEP.

La derivazione delle equazioni di inferenza percettiva dalla minimizzazione dell'energia libera richiede di specificare la forma del modello generativo. Per un modello lineare gerarchico — la classe più studiata in connessione con il predictive coding corticale — il modello generativo ha la forma:

$$s = g^1(\mu^1) + z^1, z^1 \sim \mathcal{N}(0, \Pi^{1-1}) \quad \mu^1 = g^2(\mu^2) + z^2, z^2 \sim \mathcal{N}(0, \Pi^{2-1})$$

dove g^1 e g^2 sono funzioni generative (tipicamente lineari nei modelli di Friston originali, non lineari nelle formulazioni più recenti), μ^1 e μ^2 sono le medie posteriori delle cause latenti ai livelli 1 e 2, e Π^1 , Π^2 sono le matrici di precisione (inverse delle matrici di covarianza) che pesano l'importanza relativa degli errori di predizione ai diversi livelli. Per questo modello, l'energia libera variazionale è:

$$F = \frac{1}{2} \epsilon^1 T \Pi^1 \epsilon^1 + \frac{1}{2} \epsilon^2 T \Pi^2 \epsilon^2 - \frac{1}{2} \log |\Pi^1| - \frac{1}{2} \log |\Pi^2| + \text{const}$$

dove $\epsilon^1 = s - g^1(\mu^1)$ è l'errore di predizione al livello 1 e $\epsilon^2 = \mu^1 - g^2(\mu^2)$ è l'errore di predizione al livello 2 (la discrepanza tra la stima di livello 1 e la predizione generata dal livello 2). La minimizzazione di F rispetto alle medie posteriori μ^1 produce:

$$\mu^1 = -\partial F / \partial \mu^1 = \Pi^1 \epsilon^1 \cdot (\partial g^1 / \partial \mu^1) - \Pi^2 \epsilon^2$$

Il primo termine spinge μ^1 verso la direzione che riduce l'errore di predizione sensoriale (accuratezza della spiegazione), mentre il secondo termine spinge μ^1 verso la direzione che riduce la discrepanza con la predizione del livello superiore (coerenza gerarchica).

La minimizzazione di F rispetto alle medie posteriori μ^2 produce:

$$\mu^2 = -\partial F / \partial \mu^2 = \Pi^2 \epsilon^2 \cdot (\partial g^2 / \partial \mu^2)$$

dove l'unico termine è la pressione verso la riduzione dell'errore di predizione al livello 2. Questo schema di aggiornamento — minimizzazione locale dell'energia libera attraverso errori di predizione pesati con la precisione — è la realizzazione formale del predictive coding nel framework del FEP.

7.3 Active Inference: dall'inferenza percettiva all'azione

Una delle estensioni più significative del FEP è l'Active Inference, in cui l'azione viene ricondotta allo stesso principio di minimizzazione dell'energia libera che governa la percezione. Il salto concettuale è notevole: nell'Active Inference, l'agente non massimizza una funzione di ricompensa esterna (come nel reinforcement learning classico) ma minimizza la propria energia libera attraverso una combinazione di inferenza (aggiornamento delle credenze per spiegare meglio le osservazioni) e di azione (modifica dell'ambiente per renderlo conforme alle aspettative).

Il meccanismo formale dell'azione nell'Active Inference è il seguente. Le azioni a_t vengono selezionate per minimizzare l'energia libera rispetto alle azioni stesse, ovvero per ridurre la discrepanza tra lo stato sensoriale atteso e lo stato sensoriale attuale:

$$\dot{a} = -\partial F / \partial a = -\partial F / \partial s \cdot \partial s / \partial a$$

dove $\partial s / \partial a$ è il jacobiano degli stati sensoriali rispetto alle azioni (come le azioni cambiano le osservazioni). Nella formulazione più intuitiva, l'agente mantiene credenze sullo stato futuro desiderato — non una funzione di ricompensa, ma preferenze sensoriali codificate nel prior $p(s)$ del modello generativo — e seleziona azioni che cambiano il mondo per renderlo più conforme alle proprie preferenze.

Questa formulazione unifica percezione e azione sotto lo stesso principio variazionale, con una struttura che è formalmente molto simile a quella della pianificazione model-based di LeCun. La differenza principale è di enfasi: nel FEP l'agente minimizza la sorpresa (si aspetta di stare in stati preferiti e agisce per realizzare queste aspettative), nel framework di LeCun il pianificatore minimizza una funzione di costo esplicita (si cerca di raggiungere stati obiettivo attraverso azioni pianificate). Queste due formulazioni sono formalmente equivalenti quando il costo lecuniano è identificato con l'energia libera fristoniana e il modulo di costo con il prior sugli stati preferiti — ma la corrispondenza non è perfetta perché LeCun prevede componenti di costo acquisite che non hanno un naturale equivalente nel FEP.

Il FEP viene esteso da Friston e collaboratori (Friston et al., 2017) alla pianificazione attraverso l'introduzione delle credenze sulle politiche π (sequenze di azioni future). Nella formulazione dell'Active Inference estesa a sequenze temporali, l'agente mantiene credenze $Q(\pi)$ sulla distribuzione delle politiche e le aggiorna minimizzando l'energia libera attesa (Expected Free Energy, EFE) di ciascuna politica:

$$G(\pi) = \sum_{\tau} \{ \mathbb{E}_{-} \{ Q(s_{-\tau}, o_{-\tau} \mid \pi) \} [\log Q(s_{-\tau} \mid \pi) - \log p(s_{-\tau}, o_{-\tau})] \}$$

dove $s_{-\tau}$ è lo stato al tempo futuro τ e $o_{-\tau}$ è l'osservazione. Espandendo questo termine:

$$G(\pi) = \sum_{\tau} \{ -\mathbb{E}_{-} \{ Q(o_{-\tau} \mid \pi) \} [\log p(o_{-\tau})] - \mathbb{E}_{-} \{ Q(s_{-\tau} \mid \pi) \} [H(P(o_{-\tau} \mid s_{-\tau}))] \}$$

Il primo termine è la sorpresa attesa sull'osservazione — quanto le osservazioni future attese sotto π differiscono dalle preferenze dell'agente (codificate in $p(o_{-\tau})$). Il secondo termine è l'incertezza attesa sull'osservazione condizionatamente allo stato — misura quanto le azioni selezionate dalla politica π ridurrebbero l'incertezza epistémica dell'agente sul mondo. Politiche con bassa EFE combinano dunque il perseguimento di stati preferiti (riduzione della sorpresa pragmatica) con l'esplorazione informativa (riduzione dell'incertezza epistémica). Questa forma di equilibrio esplorazione-sfruttamento emerge naturalmente dalla struttura matematica dell'EFE, senza necessità di meccanismi aggiuntivi come l' ϵ -greedy exploration o UCB.

7.4 Markov Blankets e la struttura dell'identità del sistema

Un aspetto del FEP che ha implicazioni profonde per la concezione di autonomia e identità dei sistemi intelligenti è il concetto di Markov blanket, introdotto da Judea Pearl (1988) nella teoria delle reti bayesiane e ripreso e generalizzato da Friston (2013) nel contesto del FEP.

Nel contesto di un sistema dinamico, il Markov blanket di un insieme di variabili interni μ è l'insieme minimale di variabili esterne che rendono μ condizionalmente indipendente dal resto del sistema. Formalmente, in una rete bayesiana, il Markov blanket $MB(\mu)$ di un nodo μ è costituito dai suoi genitori (cause dirette), figli (effetti diretti) e co-genitori (altre cause dei figli di μ). La proprietà chiave è:

$$P(\mu \mid MB(\mu)) = P(\mu \mid \text{tutto il resto})$$

ovvero, condizionatamente al suo Markov blanket, μ è indipendente da tutte le altre variabili del sistema.

Friston argomenta che qualunque sistema auto-organizzato deve possedere un Markov blanket: senza una separazione statistica tra interno ed esterno, il sistema non avrebbe un'identità definita, si dissolverebbe nell'ambiente. In un agente cognitivo, il Markov blanket si compone di stati attivi (le azioni che l'agente esercita sull'ambiente) e stati sensoriali (gli stati dell'ambiente che influenzano le percezioni dell'agente). Gli stati interni (μ , le credenze del cervello) sono condizionalmente indipendenti dall'ambiente dati gli stati del Markov blanket.

La rilevanza di questo concetto per la proposta di LeCun è duplice. Primo, fornisce un fondamento matematico per la separazione tra l'agente e il suo ambiente — la distinzione tra "dentro" e "fuori" che è prerequisito concettuale per qualunque teoria dell'agentività. Secondo, suggerisce che l'architettura modulare di LeCun — con il modulo percettivo come interfaccia sensoriale, il modulo attoriale come interfaccia motoria, e il modello di mondo come stato interno — è formalmente isomorfa alla struttura MB-agente-ambiente del FEP.

7.5 Connessioni e divergenze tra JEPA e il Free Energy Principle

Le analogie tra JEPA e il FEP sono profonde e non accidentali: entrambi i framework enfatizzano la predizione come principio organizzativo della cognizione, entrambi trattano la percezione come inferenza piuttosto che come estrazione passiva di feature, entrambi integrano apprendimento e pianificazione in un unico framework, entrambi trattano l'incertezza come variabile chiave da modellare esplicitamente. Tuttavia, le differenze sono altrettanto significative e vale la pena articolare con precisione.

La prima differenza riguarda il livello di analisi e l'ambizione normativa. Il FEP è un framework normativo — descrive il principio unico che, secondo Friston, governa tutti i sistemi auto-organizzati, dal singolo neurone all'intera corteccia, dall'ameba all'essere umano. JEPA è un framework ingegneristico — propone un'architettura concreta implementabile su hardware disponibile per un compito specifico (l'apprendimento autosupervisionato di rappresentazioni del mondo fisico). Il FEP non dice come implementare la minimizzazione dell'energia libera in un sistema neurale o artificiale concreto; JEPA non pretende di essere il principio universale della cognizione.

La seconda differenza riguarda il trattamento della precisione (precision). Nel FEP, la precisione — l'inverso della varianza degli errori di predizione — è un parametro fondamentale che regola il peso relativo degli errori di predizione ai diversi livelli della gerarchia e in diverse condizioni. La regolazione della precisione è il meccanismo attraverso cui il FEP spiega fenomeni come l'attenzione (amplificazione degli errori di predizione nelle regioni di interesse), la coscienza (alti livelli di precisione nelle rappresentazioni accessibili globalmente) e alcune sindromi psicopatologiche come la schizofrenia (aberrazioni nei meccanismi di regolazione della precisione che producono errori di predizione "fantasma" non ancorati a stimoli reali). JEPA non ha un meccanismo equivalente di regolazione dinamica della precisione; le sue perdite di addestramento hanno pesi fissi, non adattativi.

La terza differenza riguarda il tipo di variabili latenti. Nel FEP, le variabili latenti z sono interpretate come le cause fisiche, biologicamente significative, degli stati sensoriali — entità come "luce da sinistra", "volto a distanza 50 cm", "voce amica a 3 metri". In JEPA, le variabili latenti z che codificano l'incertezza irriducibile non hanno una tale interpretazione esplicita: sono vettori numerici il cui significato è definito implicitamente dalla struttura dell'addestramento. Questa differenza di interpretazione ha conseguenze per la trasparenza e l'analizzabilità del sistema: un modello nel senso del FEP è, in linea di principio, interpretabile; le variabili latenti di JEPA possono

essere visualizzate e analizzate ma non sono necessariamente interpretabili in termini fisici intuitivi.

La quarta differenza riguarda l'Active Inference e il rapporto tra percezione e azione. Nel FEP, percezione e azione sono aspetti dello stesso processo di minimizzazione dell'energia libera: l'agente riduce l'energia libera sia aggiornando le credenze (inferenza) sia agendo sul mondo (active inference). In JEPA nella sua formulazione attuale, il modello del mondo e il pianificatore sono componenti separati con ruoli distinti; non è chiaro come la predizione autosupervisionata del modello del mondo sia connessa, durante l'addestramento, alle azioni del sistema. La proposta lecuniana non ha ancora un equivalente dell'active inference fristoniana — un meccanismo che integri l'apprendimento del modello di mondo con l'esperienza dell'agente che agisce nel mondo.

7.6 Evidenze neuroscientifiche per il predictive coding corticale

La solidità empirica del predictive coding come modello della corteccia cerebrale è importante per valutare la plausibilità biologica della proposta di LeCun, che si ispira esplicitamente alle architetture cognitive biologiche. Le evidenze sono molteplici, eterogenee, e non conclusive — come è normale per qualunque teoria della cognizione.

Le evidenze più dirette provengono da studi di neuroimaging e neurofisiologia che documentano il soppresso di risposta ai stimoli attesi e l'amplificazione di risposta agli stimoli anomali (Prediction Error Responses, PER). Gli studi di fMRI mostrano che le aree corticali primarie — V1 per la visione, A1 per l'udito — attivano in modo più robusto agli stimoli inattesi rispetto agli stimoli attesi, anche quando l'intensità dello stimolo fisico è identica. Questo pattern — Mismatch Negativity (MMN) nell'EEG, amplificazione BOLD in fMRI — è esattamente ciò che predice un sistema di predictive coding che trasmette errori di predizione piuttosto che segnali sensoriali interi.

Le evidenze da studi di lesione e farmacologia sono più indirette ma converging. Pazienti con lesioni prefrontali — che nel quadro gerarchico del predictive coding corrispondono ai livelli più alti del modello generativo — mostrano pattern di difficoltà nei compiti che richiedono la generazione di predizioni sulla base di contesti precedenti (working memory, attenzione selettiva, pianificazione sequenziale). La farmacologia della dopamina nel sistema cortico-striatale è stata interpretata da Friston e collaboratori come regolazione della precisione delle predizioni: la dopamina aumenta il guadagno degli errori di predizione nelle aree striatali, amplificando la segnalazione di eventi inaspettati che hanno valore per l'apprendimento.

Più direttamente rilevante per la proposta di LeCun sono le evidenze sulla struttura gerarchica delle rappresentazioni corticali visive. Studi di neuroimaging (Kay et al., 2008; Huth et al., 2012) hanno mostrato che le aree visive di alto livello (aree occipito-temporali, aree fusiformi) codificano proprietà semantiche degli stimoli visivi — la categoria dell'oggetto, il suo significato funzionale, le sue relazioni causali con

altri oggetti — in modo che non è riducibile alle proprietà sensoriali di basso livello. Questo è coerente con l'idea che la gerarchia corticale visiva realizzi una forma di astrazione progressiva analoga a quella delle reti neurali convoluzionali profonde — un'analogia che LeCun cita esplicitamente come motivazione del design gerarchico del modulo percettivo in JEPA.

7.7 Apprendimento per osservazione, sviluppo infantile e "core knowledge"

L'argomento che LeCun considera più convincente per giustificare l'approccio autosupervisionato è il paragone con lo sviluppo cognitivo infantile. I bambini, argomenta LeCun, costruiscono una comprensione sofisticata della fisica degli oggetti, delle relazioni causali e della struttura dello spazio attraverso osservazione passiva e interazione spontanea, con una frugalità di dati che i modelli di apprendimento artificiale non avvicinano nemmeno lontanamente.

Questa intuizione è supportata da un corpus sostanziale di evidenze sperimentali sulla psicologia dello sviluppo. I lavori di Renée Baillargeon, sviluppati nel corso degli anni Ottanta e Novanta attraverso il paradigma della violazione dell'aspettazione, hanno dimostrato che bambini di 2.5–3.5 mesi di vita mostrano già una comprensione rudimentale della permanenza dell'oggetto (gli oggetti continuano ad esistere quando non sono visibili), della solidità (gli oggetti fisici non possono attraversarsi) e della coerenza spaziale (gli oggetti si muovono in traiettorie continue). Il paradigma sfrutta il fatto che i bambini guardano più a lungo gli eventi "impossibili" — violazioni delle aspettative fisiche — rispetto agli eventi attesi, permettendo di inferire le aspettative senza richiedere risposta motoria.

Elizabeth Spelke e collaboratori (Spelke, 1990; Spelke e Kinzler, 2007) hanno proposto la teoria del "core knowledge" per descrivere queste conoscenze innate o precoci: un insieme di sistemi cognitivi dominio-specifici che forniscono rappresentazioni di oggetti fisici, agenti intenzionali, numeri e geometria che sono disponibili dalla nascita o si sviluppano molto precocemente attraverso l'interazione con l'ambiente. Questi sistemi di core knowledge sono automatici, robusti e si manifestano in modo simile in culture diverse — suggerendo una base biologicamente preformata piuttosto che una costruzione puramente culturale.

Karen Wynn (1992), in un esperimento divenuto classico, ha dimostrato che bambini di cinque mesi mostrano sorpresa (guardano più a lungo) quando il risultato di un'operazione di addizione o sottrazione visiva di oggetti viola le aspettative aritmetiche: un display che mostra $1+1 = 1$ o $1+1 = 3$ produce più attenzione di un display che mostra $1+1 = 2$. Questo risultato, replicato e raffinato in numerosi laboratori, suggerisce che la capacità di tracking numerale approssimativo abbia radici evolutive profonde.

La rilevanza di queste ricerche per la proposta di LeCun è nel principio che le strutture di conoscenza più fondamentali non vengono acquisite attraverso istruzione esplicita o rinforzo sistematico, ma emergono dall'interazione con il mondo fisico in modo che il

sistema cognitivo è biologicamente attrezzato per supportare. JEPA aspira a implementare qualcosa di analogo: fornire all'architettura artificiale i principi organizzativi (predizione in spazio latente, struttura gerarchica, modulazione dell'incertezza attraverso variabili latenti) che le permettano di sviluppare strutture di conoscenza del mondo fisico attraverso osservazione autosupervisionata, analogamente a come la gerarchia corticale biologica sviluppa le proprie strutture rappresentazionali attraverso l'esperienza sensoriale dei primi anni di vita.

Il punto di tensione in questo paragone è che il sistema cognitivo biologico porta con sé, alla nascita, un'enorme quantità di struttura innata — un inductive bias evolutivamente ottimizzato che ha richiesto milioni di anni di selezione per emergere. JEPA porta con sé l'architettura (predittore in spazio latente, encoder con momentum, mascheramento strutturato) ma non i bias evolutivi profondi che permettono ai bambini di acquisire conoscenze fisiche di base così rapidamente. Questa discrepanza tra il potere espressivo dell'architettura artificiale e gli inductive bias biologici pre-specificati è uno dei punti critici più rilevanti della proposta lecuniana, che verrà approfondito nella sezione dedicata alla valutazione critica.

7.8 Common sense fisico come struttura causale implicita

LeCun usa il termine "common sense" in un senso tecnico che merita di essere precisato, per evitare la confusione con i significati più comuni dell'espressione. Non intende l'insieme delle credenze culturalmente condivise né la capacità di navigare situazioni sociali ambigue. Intende la conoscenza implicita, non verbalizzata, della fisica degli oggetti, della geometria dello spazio, delle relazioni causali meccaniche — la conoscenza che permette a un bambino di due anni di prevedere che un oggetto cadrà se non è supportato, che un liquido si riverserà se il contenitore viene eccessivamente inclinato, che un oggetto lanciato seguirà una traiettoria parabolica approssimativa.

Questa forma di conoscenza ha una struttura che la distingue dalla conoscenza esplicita e verbalizzabile. Primo, è procedurale anziché dichiarativa: non si manifesta come affermazioni su come funziona il mondo ma come capacità di prevedere, pianificare e agire in modo fisicamente sensato. Secondo, è altamente generalizzabile: la comprensione della gravità come principio non richiede di aver visto cadere ogni specifico tipo di oggetto. Terzo, è robusta rispetto a variazioni superficiali: la stessa comprensione fisica si applica a oggetti di dimensioni, colori, materiali e forme molto diversi.

La critica di LeCun agli LLM su questo punto è operazionalmente precisa: modelli linguistici addestrati su testo possono produrre affermazioni corrette sui principi fisici (possono scrivere che la gravità accelera gli oggetti cadenti di 9.8 m/s^2), ma mostrano pattern di fallimento sistematici in compiti di ragionamento fisico che richiedono l'applicazione di questi principi a situazioni concrete nuove. La dissociazione tra la competenza linguistica (saper parlare di fisica) e la competenza causale (saper ragionare fisicamente) è, per LeCun, la firma di un sistema che ha acquisito pattern statistici sul linguaggio della fisica senza acquisire la struttura causale sottostante.

V-JEPA, addestrato su video del mondo fisico, dovrebbe in linea teorica acquisire strutture rappresentazionali che catturano le regolarità causali di cui gli LLM mancano: la continuità delle traiettorie, la conservazione della massa, la persistenza degli oggetti nell'occlusione. La domanda empiricamente aperta è se le strutture appresi da V-JEPA attraverso il compito di predizione in spazio latente siano effettivamente causali nel senso richiesto — se catturino i principi fisici generativi o se apprendano correlazioni statistiche sofisticate che appaiono causali nei casi osservati ma collassano in distribuzione out-of-training. Questa è la grande verifica empirica che il programma di ricerca leCuniano deve ancora affrontare sistematicamente.

8. Confronto con Paradigmi Concorrenti

8.1 La Scaling Hypothesis: fondamenti formali e limiti strutturali

La scaling hypothesis — l'idea che la performance dei modelli linguistici migliori in modo prevedibile e monotono al crescere del numero di parametri, dei dati di addestramento e della potenza di calcolo — è diventata, nel corso degli ultimi anni, qualcosa di più di un'ipotesi empirica: è diventata una strategia industriale, un programma di ricerca, quasi una visione del mondo. Comprendere la critica di LeCun a questa strategia richiede di esaminare dapprima il formalismo matematico che la sostiene, poi le sue implicazioni logiche, e infine le condizioni al contorno che ne delimitano la validità.

Il lavoro fondativo è quello di Kaplan e collaboratori (2020), che ha stabilito leggi di scala (scaling laws) empiriche per i modelli linguistici Transformer. Attraverso esperimenti condotti su modelli che variano da 10^7 a 10^{10} parametri e su corpus che variano da 10^8 a 10^{11} token, Kaplan et al. hanno trovato che la perdita di test (cross-entropy su un corpus di validazione) segue una legge di potenza in funzione del numero di parametri N e del numero di token di addestramento D :

$$L(N) \approx (N_c/N)^{\alpha_N} + L_{\text{irr}}, \text{ per } D \rightarrow \infty \quad L(D) \approx (D_c/D)^{\alpha_D} + L_{\text{irr}}, \text{ per } N \rightarrow \infty$$

dove $N_c \approx 8.8 \times 10^{13}$, $\alpha_N \approx 0.076$, $D_c \approx 5.4 \times 10^{13}$, $\alpha_D \approx 0.095$, e $L_{\text{irr}} \approx 1.69$ sono costanti empiriche determinate dal fit sui dati sperimentali. Il modello congiunto è:

$$L(N, D) \approx (N_c/N)^{\alpha_N} + (D_c/D)^{\alpha_D} + L_{\text{irr}}$$

Queste leggi di scala hanno implicazioni pratiche immediatamente operative: dato un budget computazionale fisso C (misurato in FLOPs — floating point operations), il numero ottimale di parametri N_{opt} e di token D_{opt} che minimizzano $L(N, D)$ può essere determinato analiticamente. La relazione tra compute C , numero di parametri N e numero di token D per un singolo passo di addestramento è:

$$C \approx 6 \cdot N \cdot D \text{ (in FLOPs, approssimazione standard per Transformer densi)}$$

Minimizzare $L(N,D)$ sotto il vincolo $C = 6 \cdot N \cdot D$ attraverso moltiplicatori di Lagrange produce:

$$\partial L / \partial N = \partial L / \partial D \quad dN/dD \mid_{\{C \text{ costante}\}}$$

$$\text{ovvero: } -\alpha_N \cdot N^{\alpha_N-1} \cdot D^{\alpha_D} = -\alpha_D \cdot N^{\alpha_N} \cdot D^{\alpha_D-1} \cdot (-D/N)$$

che semplifica in: $\alpha_N/N = \alpha_D/D \cdot (D/N)$, ovvero $\alpha_N = \alpha_D$ — una condizione soddisfatta approssimativamente dalle stime empiriche di Kaplan et al. ($\alpha_N \approx 0.076$, $\alpha_D \approx 0.095$). Hoffmann e collaboratori (2022) hanno ricalibrato queste stime con dati più precisi e su scale più ampie, trovando $N_{\text{opt}} \propto C^{0.5}$ e $D_{\text{opt}} \propto C^{0.5}$ come relazione ottimale — ovvero il numero di token deve scalare proporzionalmente al numero di parametri, un risultato che contraddiceva la pratica degli anni precedenti (GPT-3, Gopher) in cui i modelli erano significativamente sotto-addestrati rispetto alla loro dimensione.

Nella formulazione più ambiziosa della scaling hypothesis — quella che LeCun attacca — si assume che queste leggi di scala si protraggano indefinitamente: per $N, D \rightarrow \infty$ e $C \rightarrow \infty$, $L(N,D) \rightarrow L_{\text{irr}}$, l'entropia irriducibile dei dati, che rappresenta il limite inferiore assoluto. E si assume, ulteriormente, che al tendere di L verso L_{irr} , le capacità cognitive dei modelli si avvicinino all'intelligenza generale.

La critica di LeCun si articola a questo livello. Il problema non è la validità empirica delle leggi di scala nel loro regime osservato — sono robuste, ben documentate, e hanno guidato previsioni accurate. Il problema è la terza assunzione: che la riduzione di L implichi l'acquisizione delle capacità cognitive rilevanti per l'intelligenza generale. Questa assunzione non è implicata dalle leggi di scala stesse: le leggi di scala descrivono come L dipende da N e D , non cosa rappresenta L rispetto alle capacità cognitive.

L'argomento formale può essere strutturato come segue. La perdita di test L misura:

$$L(N,D) = H(P_{\text{data}}, P_{\theta}) = H(P_{\text{data}}) + KL(P_{\text{data}} \parallel P_{\theta})$$

dove P_{data} è la distribuzione empirica dei token nel corpus e P_{θ} è la distribuzione del modello. Minimizzare L equivale a minimizzare $KL(P_{\text{data}} \parallel P_{\theta})$, ovvero a fare convergere P_{θ} verso P_{data} . Ma se P_{data} è la distribuzione dei token nel corpus di testo, P_{data} riflette la struttura statistica del testo — le regolarità linguistiche, sintattiche, semantiche e pragmatiche del linguaggio scritto — non la struttura causale del mondo fisico. Un modello che raggiunge $L = L_{\text{irr}}$ ha appreso perfettamente la distribuzione statistica dei token di testo; non ha appreso necessariamente la fisica degli oggetti, il ragionamento causale controfattuale, o la pianificazione multi-step in ambienti fisici inediti, se queste capacità non sono pienamente codificate nella distribuzione P_{data} .

LeCun formalizza questo argomento attraverso la nozione di "informazione causale" contrapposta all'"informazione statistica". La distribuzione P_{data} è generata da agenti umani incorporati che agiscono nel mondo fisico e comunicano attraverso il

linguaggio; in questo senso, il testo contiene informazioni indirette sulla struttura causale del mondo. Ma queste informazioni sono mediate, compresse e distorte dal processo di traduzione dell'esperienza fisica in linguaggio — un processo che necessariamente perde l'informazione causale strutturale che non è verbalizzabile. La quantità di informazione causale codificata in P_data è enormemente inferiore alla quantità di informazione causale disponibile nell'esperienza sensorimotoria diretta del mondo fisico.

Questa critica ha una risposta potenziale: i modelli multimodali, addestrati su coppie testo-immagine, testo-video, testo-audio, accedono a P_data distribuzioni che contengono informazione sensoriale diretta oltre all'informazione linguistica. LeCun riconosce questo fatto ma mantiene che l'allineamento tra testo e dati sensoriali non produce strutture rappresentazionali causalmente robuste — l'addestramento continua a essere ottimizzato per la verosimiglianza del prossimo token (o dell'immagine condizionatamente al testo), non per la predizione delle strutture causali invarianti del mondo fisico. La distinzione tra modalità di addestramento e qualità delle strutture rappresentazionali apprese è al cuore del disaccordo.

8.2 L'approccio neuro-simbolico: architetture ibride e ragionamento strutturato

L'approccio neuro-simbolico costituisce, nel panorama delle alternative a LLM e JEPa, la proposta forse più ambiziosa di riconciliare i due grandi paradigmi dell'intelligenza artificiale del XX secolo: il paradigma connessionista (reti neurali, apprendimento da dati) e il paradigma simbolico (sistemi basati su regole, logica formale, rappresentazioni strutturate). La sua storia è lunga quanto quella dell'AI stessa — le prime proposte di integrazione risalgono agli anni Sessanta — e le sue realizzazioni recenti sono tecnicamente sofisticate.

Il problema che l'approccio neuro-simbolico cerca di risolvere è preciso: i sistemi neurali sono eccellenti nell'apprendimento di pattern da dati ma fragili nella generalizzazione sistematica a combinazioni non viste di concetti; i sistemi simbolici sono eccellenti nella generalizzazione sistematica ma richiedono conoscenza codificata manualmente e sono fragili rispetto a input fuori distribuzione. L'obiettivo è combinare la robustezza adattiva dei sistemi neurali con la precisione e la trasparenza dei sistemi simbolici.

Le realizzazioni più avanzate di questo programma includono Neural Theorem Provers (Evans e Grefenstette, 2018), che imparano a costruire prove logiche attraverso operazioni differenziabili su rappresentazioni neurali di fatti e regole; AlphaGeometry (Trinh et al., 2024), che combina un modello linguistico per la generazione di costruzioni geometriche ausiliarie con un dimostratore simbolico per la verifica formale delle prove; e gli approcci Program Synthesis che usano modelli neurali per generare programmi simbolici eseguibili da dati di addestramento.

Uno dei framework teorici più influenti nell'approccio neuro-simbolico recente è la Probabilistic Soft Logic (PSL) e le sue estensioni neurali. Nella PSL, le regole logiche sono formulate come vincoli di disuguaglianza a valore in $[0,1]$ anziché in $\{0,1\}$:

$$\forall x, y: \text{Amico}(x,y) \wedge \text{Voti}(y,p) \rightarrow \text{Voti}(x,p)$$

viene tradotta in:

$$\text{Voti}(x,p) \geq \text{Amico}(x,y) + \text{Voti}(y,p) - 1$$

e il problema di inferenza diventa:

$$\min_{\mathbf{v} \in [0,1]^n} \sum_r w_r \cdot \max(0, w_v(r))$$

dove w_r è il peso della regola r , $\mathbf{v}$ è il vettore dei valori di verità delle variabili, e $w_v(r)$ è la violazione della regola r . Questo problema di ottimizzazione è convesso e risolvibile in modo efficiente. L'integrazione con sistemi neurali avviene attraverso la derivazione dei pesi w_r da reti neurali addestrate su dati.

Il framework NSCL (Neuro-Symbolic Concept Learner, Mao et al., 2019) integra reti neurali per la percezione visiva con un interprete simbolico eseguibile. Dati un'immagine e una domanda in linguaggio naturale (Visual Question Answering, VQA), NSCL: (1) usa una CNN per estrarre rappresentazioni degli oggetti nell'immagine, (2) usa un modulo linguistico per analizzare la domanda e produrre un programma simbolico, (3) esegue il programma simbolico sulla rappresentazione dell'immagine per produrre la risposta. Il programma simbolico è un albero di operazioni che manipolano insiemi di oggetti: `filter_color(red)` seleziona gli oggetti rossi, `relate(left_of, x)` seleziona gli oggetti a sinistra di x , `count()` conta gli elementi di un insieme.

La posizione di LeCun rispetto all'approccio neuro-simbolico è complessa e merita una trattazione che eviti sia la caricatura della critica sia la banalizzazione dei punti di convergenza. LeCun condivide con i proponenti dell'approccio neuro-simbolico la convinzione che le rappresentazioni strutturate e causali siano necessarie per l'intelligenza genuina, e che i sistemi puramente sub-simbolici non le sviluppino spontaneamente in modo sufficiente. Ma dissente sulla terapia: ritiene che le strutture debbano emergere dall'apprendimento dai dati piuttosto che essere imposte dall'ingegnere attraverso linguaggi di rappresentazione manualmente progettati.

Il nucleo dell'obiezione lecuniana ai sistemi simbolici è quello che ha articolato costantemente fin dagli anni Novanta: la fragilità rispetto all'acquisizione. Codificare la conoscenza del mondo in strutture simboliche richiede l'enorme sforzo dell'ingegneria della conoscenza — esperti che definiscono ontologie, relazioni, regole — e produce sistemi che funzionano eccellentemente nel dominio per cui sono stati costruiti ma collassano alla prima deviazione dalla distribuzione per cui la conoscenza è stata specificata. Il limite di questa critica, nel contesto degli approcci neuro-simbolici recenti, è che sistemi come AlphaGeometry non richiedono codifica manuale della conoscenza geometrica: apprendono da dati. L'obiezione lecuniana deve essere

riformulata: i sistemi che combinano apprendimento neurale con esecuzione simbolica richiedono di specificare a priori il linguaggio del programma simbolico (il tipo di operazioni eseguibili, la struttura dell'albero di programma), il che introduce un inductive bias esplicito che potrebbe non essere generalizzabile a domini diversi da quelli per cui è stato progettato.

La questione aperta è se un sistema che apprende da osservazione del mondo fisico — come aspira JEPA — possa sviluppare strutture rappresentazionali sufficientemente strutturate da supportare ragionamento compositazionale robusto, o se la struttura compositazionale richieda un'architettura che includa componenti simbolici espliciti. Questa è, in ultima analisi, una questione empirica che il programma di ricerca lecuniano deve ancora affrontare in modo definitivo.

8.3 Il Neuro-Symbolic AI di Yoshua Bengio: dalla critica al programma alternativo

Yoshua Bengio — che condivide con LeCun il Premio Turing 2018 per i contributi fondativi al deep learning — ha negli ultimi anni abbracciato posizioni che convergono parzialmente con quelle di LeCun sulla necessità di andare oltre gli LLM, ma divergono significativamente sulla strategia. La sua proposta si concentra sull'apprendimento di rappresentazioni disentangled (disentangled representations) e sulla distinzione tra "System 1" (ragionamento veloce, intuitivo, statistico) e "System 2" (ragionamento lento, deliberativo, causale), ispirata alla psicologia cognitiva di Kahneman.

Nella prospettiva di Bengio, la maggior parte dei sistemi di deep learning attuali — incluso JEPA — realizza solo il System 1: l'apprendimento di pattern statistici veloci che permettono risposte rapide a situazioni familiari. Il System 2, necessario per il ragionamento causale deliberativo, richiede strutture rappresentazionali qualitativamente diverse: rappresentazioni sparse e disentangled in cui ciascuna dimensione corrisponde a un fattore causale interpretabile del mondo.

Il framework teorico di Bengio per le disentangled representations è formalmente collegato alla causal representation learning di Schölkopf e collaboratori. In un sistema di variabili causali $(X_1, X_2, \dots, X_k)$, dove $X_i \rightarrow X_j$ indica che X_i è causa diretta di X_j , le rappresentazioni causali sono tali che:

$$p(X_1, \dots, X_k) = \prod_i p(X_i \mid \text{PA}(X_i))$$

dove $\text{PA}(X_i)$ sono i genitori di X_i nel grafo causale. Le variabili X_i non sono i dati osservati ma le loro cause latenti — le entità fisiche, le loro proprietà, le loro relazioni. Un sistema che apprende rappresentazioni causali è un sistema le cui rappresentazioni latenti sono isomorfe a questo grafo causale: ogni dimensione della rappresentazione corrisponde a un singolo fattore causale, e le relazioni tra dimensioni corrispondono alle relazioni causali tra fattori.

La causal representation learning si ricollega al programma di Judea Pearl sulla struttura del ragionamento causale attraverso i do-calculus. Il do-calculus fornisce un

formalismo per distinguere l'osservazione (vedere che $X = x$ accade) dall'intervento (fare in modo che $X = x$): $p(Y | X = x)$ vs $p(Y | \text{do}(X = x))$. Il ragionamento causale richiede il calcolo di distribuzioni di intervento, che non sono in generale recuperabili dalla sola distribuzione osservazionale $p(X, Y)$ senza conoscenza della struttura causale.

L'implicazione per i sistemi di apprendimento è che un sistema che apprende solo dalla distribuzione osservazionale dei dati — come fanno tutti i modelli di apprendimento profondo attuali, incluso JEPA — non può in linea di principio acquisire strutture causali che richiedono la distinzione tra correlazione e causalità. Questa critica, se corretta, si applica anche a JEPA: il compito di predizione autosupervisionata apprende correlazioni tra stati del mondo (il frame precedente prevede il frame successivo), non interventi sul mondo (come cambia il futuro se si esegue un'azione specifica). La distinzione tra osservazione e intervento richiede, secondo Pearl, esperienze attive in cui il sistema manipola il mondo e osserva le conseguenze.

LeCun risponde implicitamente a questa critica attraverso il modulo attoriale e il pianificatore: l'agente nell'architettura lecuniana non è solo un osservatore passivo ma un agente che interviene e osserva le conseguenze. Il Modello di Mondo viene addestrato, nelle implementazioni complete della proposta, non solo su dati video passivi ma anche su dati di interazione — traiettorie in cui un agente robotico esegue azioni e osserva i loro effetti. Questa distinzione tra apprendimento passivo (I-JEPA, V-JEPA) e apprendimento attivo (l'architettura completa del paper del 2022) è importante: i risultati empirici esistenti riguardano il primo, mentre le ambizioni più grandi riguardano il secondo.

8.4 Active Inference come framework alternativo per l'agentività: confronto formale

L'Active Inference di Friston, discussa nel capitolo precedente dal punto di vista neuroscientifico, merita un confronto formale con il framework di pianificazione lecuniano che ne evidenzia analogie e differenze strutturali.

La struttura comune è la seguente. In entrambi i framework, un agente mantiene un modello interno del mondo (il modello generativo nel FEP, il Modello di Mondo in LeCun), usa questo modello per fare predizioni sugli stati futuri, e seleziona azioni sulla base di una funzione scalare che valuta la qualità degli stati futuri predetti (l'energia libera attesa nel FEP, il costo accumulato in LeCun). La pianificazione in entrambi i casi avviene attraverso un processo di ottimizzazione che considera sequenze di azioni future.

Le differenze formali sono, tuttavia, sostanziali. Consideriamo la funzione obiettivo della pianificazione nei due framework.

Nel framework di Active Inference con pianificazione (Friston et al., 2017), la politica ottimale π^* minimizza l'energia libera attesa (EFE):

$$\pi^* = \operatorname{argmin}_{\pi} G(\pi)$$

dove, come derivato nella sezione 7.3:

$$G(\pi) = \sum_{\tau} \{ -\mathbb{E}_{\pi} \{ Q(o_{-\tau} | \pi) \} [\log p(o_{-\tau})] - \mathbb{E}_{\pi} \{ Q(s_{-\tau} | \pi) \} [H(P(o_{-\tau} | s_{-\tau}))] \}$$

Il primo termine, $-\mathbb{E}[\log p(o_{-\tau})]$, misura la sorpresa attesa rispetto alle preferenze $p(o_{-\tau})$: più le osservazioni future attese deviamo dal prior di preferenza $p(o_{-\tau})$, più alto è il termine. Il secondo termine, $-\mathbb{E}[H(P(o_{-\tau} | s_{-\tau}))]$, misura la riduzione attesa di incertezza sugli stati futuri — un incentivo all'esplorazione informativa. La politica ottimale bilancia automaticamente queste due pressioni senza parametri aggiuntivi.

Nel framework di LeCun, il pianificatore risolve:

$$a^*\{0:T\} = \operatorname{argmin}\{a_{-\{0:T\}}\} [\sum_{t=0}^T C_{-\theta}(s_t, a_t) + C_{-T}(s_T, s_{\text{goal}})]$$

$$\text{s.t. } s_{t+1} = M_{-\theta}(s_t, a_t, z_t)$$

dove $C_{-\theta}$ è il modulo di costo e s_{goal} è lo stato obiettivo. Il costo $C_{-\theta}(s_t, a_t)$ è una funzione apprendibile che viene addestrata attraverso imitazione, rinforzo, o specifica manuale degli obiettivi intrinseci.

La differenza formale fondamentale è nella struttura delle preferenze. Nel FEP, le preferenze sono codificate nel prior $p(o_{-\tau})$ — la distribuzione sulle osservazioni future "desiderate" — che è un componente del modello generativo del sistema. Questo prior incorpora sia preferenze pragmatiche (stati del mondo che si vuole raggiungere) sia preferenze epistemiche (stati informativi che si vuole raggiungere per ridurre l'incertezza). L'EFE combina automaticamente queste due componenti senza parametri di bilanciamento aggiuntivi.

Nel framework lecuniano, le preferenze sono codificate nel modulo di costo $C_{-\theta}$ — una rete neurale addestrata separatamente attraverso supervisione o imitazione. Non c'è un meccanismo intrinseco che bilanci l'esplorazione (riduzione dell'incertezza) con lo sfruttamento (raggiungimento di stati preferiti): questo bilanciamento deve essere progettato esplicitamente dall'ingegnere, o appreso separatamente da dati. In questo senso, il framework fristoniano ha una struttura più elegante e auto-sufficiente; il framework lecuniano ha una struttura più flessibile e ingegnerizzabile.

Un'altra differenza rilevante riguarda la modellazione dell'incertezza. Nel FEP, l'incertezza è trattata in modo sistematico attraverso la propagazione delle distribuzioni probabilistiche attraverso il modello generativo — l'agente mantiene distribuzioni sugli stati latenti e sulle osservazioni future, non stime puntali. Nel framework di LeCun, le variabili latenti z_t che codificano l'incertezza sono trattate deterministicamente durante la pianificazione (si campiona z_t e si pianifica condizionatamente), e non c'è un meccanismo sistematico per propagare l'incertezza attraverso la traiettoria pianificata. Questa è una limitazione tecnica che potrebbe essere risolta attraverso l'integrazione di tecniche di inferenza probabilistica nel pianificatore lecuniano, ma non è ancora stata affrontata nell'architettura proposta.

Un tentativo di integrazione dei due framework è la lavoro di Sajid e collaboratori (2021), che riformulano l'Active Inference nel linguaggio delle reti neurali profonde, rendendo il FEP implementabile come architettura di apprendimento profondo. Questa linea di ricerca apre la possibilità di una sintesi tra il rigore matematico del FEP e la scalabilità ingegneristica degli approcci lecuniani, ma rimane ancora a un livello prevalentemente teorico.

8.5 Mixture-of-Experts e la modularità emergente nei transformer

Le architetture Mixture-of-Experts (MoE) rappresentano il tentativo più sistematico, all'interno del paradigma dei modelli linguistici di grandi dimensioni, di introdurre una forma di modularità e specializzazione funzionale che richiama, almeno superficialmente, l'architettura modulare proposta da LeCun.

Il principio fondamentale delle architetture MoE è la sostituzione della rete feed-forward densa (FFN) di ogni blocco Transformer con una collezione di N reti "esperte" $\{FFN_1, \dots, FFN_N\}$ e un meccanismo di routing che seleziona, per ogni token, quale sottoinsieme di esperti consultare. La formulazione originale di Shazeer e collaboratori (Sparsely-Gated Mixture-of-Experts, 2017) usa un router soft con top-k selection:

$$MoE(x) = \sum_{i \in \text{Top-}k(G(x))} G_i(x) \cdot FFN_i(x)$$

dove $G(x) = \text{Softmax}(x \cdot W_g)$ è il vettore di gate (rete di routing), $\text{Top-}k(v)$ seleziona gli indici dei k valori più alti del vettore v , e $G_i(x)$ è il peso normalizzato assegnato all'esperto i . Con $k \ll N$ — tipicamente $k = 2$ o $k = 4$ su $N = 64$ o $N = 128$ esperti — soltanto una piccola frazione degli esperti viene attivata per ogni token, realizzando una forma di sparse computation che aumenta la capacità del modello senza proporzionale aumento del costo computazionale per token.

La perdita di addestramento include tipicamente un termine di bilanciamento del carico (load balancing loss) che previene il fenomeno del collasso degli esperti — la tendenza del router a convergere sempre verso lo stesso sottoinsieme di esperti:

$$\mathcal{L}_{\text{balance}} = N \cdot \sum_{i=1}^N f_i \cdot p_i$$

dove $f_i = (1/T) \sum_{t=1}^T 1_{[i \in \text{Top-}k(G(x_t))]}$ è la frazione di token assegnati all'esperto i e $p_i = (1/T) \sum_{t=1}^T G_i(x_t)$ è la probabilità media di gate dell'esperto i . Questo termine penalizza distribuzioni di carico non uniformi, incoraggiando gli esperti a specializzarsi su sottodomini diversi del corpus piuttosto che convergere verso un unico esperto generico.

Modelli come Mixtral-8x7B (Jiang et al., 2024) e Gemini 1.5 Pro (Google DeepMind, 2024) hanno implementato varianti di questa architettura a scala, dimostrando che la specializzazione degli esperti emerge dall'addestramento: diversi esperti tendono a essere attivati per token di domini lessicali diversi (matematica, codice, lingue diverse, testi scientifici). Questa specializzazione emergente richiama, superficialmente, la modularità funzionale proposta da LeCun.

La critica lecuniana a questa modularità superficiale è tuttavia precisa. La specializzazione per dominio lessicale o sintattico che emerge nelle architetture MoE non corrisponde alla separazione funzionale che LeCun propone tra percezione, modello del mondo, memoria, costo e pianificatore. Gli esperti MoE sono tutti componenti della stessa funzione di stima della distribuzione linguistica; non realizzano funzioni cognitive qualitativamente diverse. La differenza è quella tra modularità di capacità (diversi esperti sono bravi su diversi tipi di testo) e modularità funzionale (componenti distinti realizzano funzioni cognitive distinte — percepire, modellare, ricordare, valutare, pianificare). L'architettura lecuniana aspira alla seconda; le MoE realizzano la prima.

Una linea di ricerca recente tenta di introdurre una modularità funzionale più profonda nei transformer attraverso meccanismi di attenzione strutturata e separazione dei ruoli delle teste di attenzione. Il lavoro di Anthropic su "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning" (Bricken et al., 2023) ha mostrato che le direzioni di attivazione all'interno dei transformer possono essere decomposte in feature monosemantiche attraverso sparse autoencoders — ovvero che le rappresentazioni interne dei transformer, pur essendo polisemantiche (ogni neurone risponde a molteplici concetti non correlati), possono essere proiettate in uno spazio di feature interpretabili e specializzate. Questo risultato suggerisce che la modularità funzionale potrebbe emergere anche all'interno di architetture non esplicitamente modulari — ma la questione se questa modularità sia funzionalmente equivalente a quella proposta da LeCun rimane aperta.

8.6 Foundation Models per la fisica: un'alternativa al World Model lecuniano?

Una linea di ricerca emersa recentemente che si colloca in modo interessante rispetto alla proposta di LeCun è lo sviluppo di foundation models specificamente addestrati per la comprensione e la simulazione della fisica. Modelli come PhysicallyGroundedLLMs (Gao et al., 2024), NeuroSymbolicPhysics (Li et al., 2023), e una serie di approcci basati su Graph Neural Networks (GNN) per la simulazione di sistemi fisici propongono di apprendere rappresentazioni della fisica degli oggetti non attraverso osservazione video autosupervisionata, ma attraverso addestramento su simulazioni fisiche con accesso a variabili di stato esplicite (masse, velocità, forze, collisioni).

L'approccio GNN per la fisica è formalmente interessante. Un sistema di N corpi interagenti può essere rappresentato come un grafo $G = (V, E)$ dove i nodi $V = \{v_1, \dots, v_N\}$ corrispondono ai corpi (con feature che includono posizione, velocità, massa) e gli archi E corrispondono alle interazioni (con feature che includono distanza, tipo di interazione, forza). La dinamica del sistema viene simulata attraverso message passing:

$$m_{\{ij\}} = \varphi_e(v_i, v_j, e_{\{ij\}}) \quad v'_i = \varphi_v(v_i, \sum_{\{j \in \mathcal{N}(i)\}} m_{\{ij\}})$$

dove φ_e è la funzione di edge update, φ_v è la funzione di node update, e $\mathcal{N}(i)$ sono i vicini del nodo i nel grafo. Questo schema — che corrisponde a un passo della variante equivariante del message passing su grafi — apprende automaticamente le leggi fisiche (le forze tra corpi, i loro effetti sulle velocità e posizioni) dai dati di traiettoria senza mai essere esplicitamente programmato con le equazioni di Newton o di Lagrange.

L'approccio di Battaglia e collaboratori in "Relational Inductive Biases, Deep Learning, and Graph Networks" (2018) è il manifesto teorico di questa linea: argumenta che i graph neural networks incorporano l'inductive bias strutturalmente adatto alla fisica — l'idea che il mondo sia composto da entità distinte che interagiscono attraverso relazioni — e che questo bias li renda più efficienti dei sistemi che trattano l'input come un tensore denso senza struttura.

In cosa differisce questo approccio da JEPA? La differenza principale è nell'inductive bias. I GNN per la fisica assumono che il mondo sia costituito da entità discrete (oggetti) con proprietà continue (posizione, velocità) che interagiscono attraverso relazioni locali (prossimità spaziale). JEPA non assume questa struttura: apprende rappresentazioni latenti senza imporre che corrispondano a oggetti discreti o a relazioni tra oggetti. L'approccio GNN ha un inductive bias più forte e più specifico — il che lo rende più efficiente su compiti di simulazione fisica ma potenzialmente meno generale su compiti che non si mappano naturalmente su sistemi di oggetti interagenti.

LeCun argumenta che le strutture di "oggetto", "relazione", "forza" non dovrebbero essere imposte dall'ingegnere ma emergere dall'apprendimento dal mondo fisico — la stessa argomentazione che muove contro l'approccio neuro-simbolico. La questione empirica è se JEPA, senza questo inductive bias esplicito, possa apprendere strutture rappresentazionali che abbiano le stesse proprietà di generalizzazione di quelle apprese dai GNN. Le evidenze attuali non sono conclusive.

8.7 Tabula Rasa vs. Inductive Biases: la questione della generalizzazione

Il dibattito tra LeCun e i proponenti di approcci neuro-simbolici, causal representation learning, e GNN ha una struttura che può essere inquadrata come tensione tra due strategie epistemologiche: la strategia "tabula rasa" (apprendiamo tutto dai dati, con la minima struttura imposta a priori) e la strategia "inductive biases" (forniamo al sistema la struttura giusta e lasciamo che l'apprendimento riempi i dettagli).

La letteratura teorica sull'apprendimento automatico suggerisce che questa tensione non è risolvibile in senso assoluto: il teorema No Free Lunch (Wolpert e Macready, 1997) stabilisce che, mediando su tutti i possibili problemi, nessun algoritmo di apprendimento è migliore di nessun altro — la superiorità di un algoritmo su una classe di problemi implica necessariamente inferiorità su un'altra classe. Ogni algoritmo che apprende efficientemente su alcuni problemi incorpora implicitamente

inductive bias — assunzioni sulla struttura dello spazio dei problemi che lo rendono più adatto a quel tipo di problema.

Formalmente, il teorema NFL nel caso dell'ottimizzazione stabilisce che, per qualunque algoritmo di ottimizzazione A e B e per qualunque funzione di performance m:

$$\sum_f m(A, f) = \sum_f m(B, f)$$

dove la somma è su tutte le possibili funzioni obiettivo f (su un dominio finito). Questo significa che la superiorità di A rispetto a B su alcune funzioni implica l'inferiorità su altre: non esiste un algoritmo "migliore in assoluto".

Nel contesto del dibattito sugli inductive bias per l'apprendimento della fisica, il teorema NFL implica che qualunque sistema che apprende la fisica dei corpi rigidi da dati deve incorporare qualche forma di bias che lo avvantaggia rispetto a sistemi senza quel bias — e che questo stesso bias lo svantaggia su problemi con struttura diversa. L'approccio GNN ha un bias verso i sistemi di entità discrete interagenti; JEPA ha un bias verso la struttura predittiva temporale del mondo fisico sensoriale. Quale bias è "meglio" dipende da quale classe di compiti si vuole risolvere.

La posizione di LeCun su questa questione è che il bias della predizione in spazio latente da dati sensoriali è il più generale possibile per un agente che deve navigare il mondo fisico: qualunque compito che un agente deve svolgere richiede la capacità di prevedere le conseguenze delle proprie azioni sul mondo sensoriale, e questa capacità è esattamente ciò che JEPA apprende. I bias più specifici (struttura a oggetti, struttura causale esplicita) sono, secondo LeCun, bias che potrebbero emergere spontaneamente da JEPA se il sistema è sufficientemente capace, e che non devono essere imposti a priori perché ciò limiterebbe la generalizzabilità del sistema a domini in cui quella struttura specifica non si manifesta.

Questa argomentazione è convincente a livello di principio ma deve ancora essere validata empiricamente: la domanda se JEPA, senza bias strutturali espliciti, possa apprendere strutture di oggetti e relazioni causali comparabili a quelle dei sistemi che le incorporano esplicitamente rimane aperta, e la sua risposta determinerà in misura significativa il successo del programma di ricerca lecuniano.

8.8 Architetture State Space Models e la concorrenza sull'efficienza computazionale

Una famiglia di architetture che ha guadagnato attenzione crescente a partire dal 2022 come alternativa ai Transformer — le State Space Models (SSM), e in particolare Mamba (Gu e Dao, 2023) — merita menzione nel confronto con la proposta di LeCun, non per la convergenza nelle motivazioni teoriche ma per la convergenza in alcune proprietà computazionali.

Le SSM sono architetture per sequenze basate su equazioni differenziali ordinarie (ODE) discrete. Il modello S4 (Gu et al., 2021) parametrizza una ODE lineare del primo ordine:

$$\dot{x}(t) = A x(t) + B u(t) \quad y(t) = C x(t) + D u(t)$$

dove $u(t)$ è il segnale di input, $x(t)$ è lo stato nascosto continuo, $y(t)$ è l'output, e A , B , C , D sono le matrici di sistema. Discretizzando questa ODE con passo Δt :

$$x_k = \bar{A} x_{k-1} + \bar{B} u_k \quad y_k = C x_k + D u_k$$

con $\bar{A} = \exp(\Delta A)$ e $\bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \Delta B$ attraverso la regola del zero-order hold. La proprietà chiave di questa parametrizzazione è l'efficienza computazionale: un SSM può essere valutato come una convoluzione causale durante l'addestramento (complessità $O(L \log L)$ per sequenze di lunghezza L) ma come un sistema ricorrente durante l'inferenza (complessità $O(1)$ per passo), combinando i vantaggi dell'addestramento parallelo dei transformer con l'efficienza computazionale dei modelli ricorrenti.

Mamba (Gu e Dao, 2023) introduce una selective state space model in cui le matrici B , C e Δ sono funzioni dell'input (input-dependent), permettendo al modello di selezionare quali informazioni mantenere nello stato in funzione del contesto — un meccanismo che realizza qualcosa di analogo all'attenzione ma con complessità lineare in L anziché quadratica.

La rilevanza di Mamba per il confronto con JEPA è limitata: le SSM sono architetture per sequenze, non architetture per modelli di mondo, e la loro motivazione principale è l'efficienza computazionale su sequenze lunghe, non la modellazione causale del mondo fisico. Tuttavia, la loro capacità di mantenere stati nascosti che evolvono nel tempo in modo selettivo sulla base dell'input suggerisce una possibile integrazione con JEPA: un modello di mondo basato su SSM potrebbe realizzare la struttura di memoria a breve termine e la dinamica temporale che LeCun propone nel suo framework con minore costo computazionale rispetto a un transformer.

Queste linee di integrazione rimangono allo stadio di speculazione teorica; le implementazioni concrete di JEPA usano fino ad oggi architetture Vision Transformer.

Version:1.0

StartHTML:0000000156

EndHTML:0000472413

StartFragment:0000000469 EndFragment:0000472405

9. I risvolti filosofici: grounding semantico, intenzionalità, coscienza, funzionalismo

La proposta di LeCun non è riducibile a un programma di ingegneria delle architetture. Essa incorpora, spesso implicitamente e talvolta esplicitamente, una posizione filosofica sulla natura delle rappresentazioni

mentali, sulla struttura della comprensione, sul rapporto tra sintassi e semantica, tra elaborazione dell'informazione e significato. Queste implicazioni filosofiche non sono un ornamento teoretico che si aggiunge a posteriori a un'impresa tecnica: esse ne costituiscono la motivazione profonda e ne delimitano le condizioni di successo. Un sistema che apprende a predire in spazio latente dal mondo fisico è, implicitamente, un tentativo di rispondere alle domande più difficili della filosofia della mente — quelle che i tecnici solitamente aggirano e i filosofi spesso formulano senza poterle testare. Il capitolo che segue non propone una ricognizione esaustiva della filosofia della mente contemporanea; propone invece un esame delle questioni che la proposta lecnuniana tocca in modo diretto e delle implicazioni che queste questioni hanno per la valutazione critica del programma.

9.1 Il problema del grounding semantico

Il problema del grounding semantico — come le rappresentazioni interne di un sistema computazionale acquistano significato rispetto al mondo che pretendono di rappresentare — è stato formulato nella sua forma più influente da Stevan Harnad nel saggio "The Symbol Grounding Problem" pubblicato su *Physica D* nel 1990. L'argomento di Harnad è elegante nella sua struttura. Consideriamo un sistema che manipola simboli — stringhe di caratteri, vettori numerici, sequenze di token — secondo regole formali. Il sistema è in grado di produrre output corretti a domande su ciò che i simboli rappresentano: se interrogato su "gatto", risponde con descrizioni accurate di gatti. Ma questa competenza è, secondo Harnad, meramente sintattica: i simboli non hanno contenuto intrinseco, sono pezzi di plastica o pattern di bit privi di riferimento. Il loro significato, per il sistema, si riduce alle relazioni con altri simboli — al fatto che "gatto" è descritto come "mammifero domestico con quattro zampe e il mantello peloso". Ma queste descrizioni sono a loro volta composte di altri simboli, altrettanto privi di contenuto intrinseco, in un regresso che non può terminare senza agganciarsi a qualcosa di non simbolico: l'esperienza diretta del mondo fisico.

La metafora che Harnad introduce per chiarire il problema è quella del dizionario monolingue per chi non conosce la lingua: ogni parola viene definita attraverso altre parole, e il sistema intero ruota su se stesso senza mai toccare il terreno. Il grounding — l'ancoraggio delle rappresentazioni simboliche a qualcosa di non simbolico — richiede, nella diagnosi di Harnad, una connessione con i processi sensorimotori attraverso cui un agente incarnato interagisce con il mondo fisico. Non è sufficiente che i simboli descrivano correttamente il mondo in termini di relazioni tra simboli: è necessario che i simboli elementari — le unità semantiche primitive — siano agganciati

direttamente a rappresentazioni iconiche e categoriali derivate dall'esperienza sensoriale.

La proposta di LeCun si iscrive, consapevolmente o meno, in una risposta al problema del grounding. I modelli linguistici autoregressivi, nella diagnosi lecuniana, non sono grounded: apprendono relazioni tra token senza mai agganciarsi al mondo fisico che quei token descrivono. JEPa, addestrato su immagini e video del mondo fisico, aspira a costruire rappresentazioni che siano ancorate all'esperienza sensoriale diretta — alla struttura ottica del mondo, alle sue regolarità causali visivamente osservabili. In questo senso, il programma JEPa può essere letto come un tentativo di risolvere il problema del grounding passando da rappresentazioni puramente linguistiche a rappresentazioni percettivo-causali.

Questa lettura, però, richiede qualificazioni importanti. Il grounding sensoriale di Harnad presuppone un agente incarnato che interagisce con il mondo — che tocca, manipola, percepisce attraverso modalità multiple in modo integrato e finalizzato. Le implementazioni attuali di JEPa — I-JEPa e V-JEPa — apprendono da immagini e video statici, senza interazione, senza propriocezione, senza feedback dall'azione. Il grounding che realizzano è, al più, un grounding percettivo parziale: le rappresentazioni sono ancorate alla struttura visiva del mondo, non alla struttura sensorimotoria dell'agire nel mondo. Harnad aveva anticipato questa distinzione distinguendo tra grounding iconico (immagini, pattern sensoriali) e grounding categoriale (categorie percettive che permettono la discriminazione sistematica): entrambi sono necessari, ma nessuno dei due è sufficiente senza l'esperienza dell'azione e delle sue conseguenze. La proposta dell'architettura completa — con il modulo attoriale e il ciclo percezione-azione — aspira a colmare questa lacuna, ma le evidenze empiriche a supporto di questa ambizione più larga sono ancora limitate.

Una complicazione ulteriore riguarda il tipo di grounding rilevante per la cognizione linguistica. Anche ammesso che JEPa realizzi un grounding percettivo robusto, rimane aperta la questione di come questo grounding si trasferisca alla comprensione del linguaggio. Il linguaggio umano non descrive solo oggetti visivamente percepibili: descrive relazioni astratte, stati interni, controfattuali, obbligo deontologico, bellezza estetica. Il grounding sensoriale è condizione necessaria ma non sufficiente per la comprensione semantica piena, il che suggerisce che la proposta lecuniana, pur affrontando il problema del grounding in modo più diretto dei sistemi puramente linguistici, non lo risolve in modo definitivo. Questo non è un difetto peculiare di JEPa: è una delle domande più difficili della filosofia del linguaggio, e nessun sistema artificiale attuale ne offre una soluzione soddisfacente.

Va registrata una tensione che percorre l'intera discussione sul grounding. Se si assume che il grounding richieda necessariamente l'esperienza soggettiva — le qualia, la prospettiva in prima persona — allora nessun sistema fisico computazionale può essere grounded nel senso rilevante, perché nessun sistema del genere ha esperienza soggettiva. Se si assume invece una definizione funzionale del grounding — le rappresentazioni sono grounded se le loro proprietà strutturali rispecchiano le proprietà strutturali del mondo che rappresentano in modo tale da supportare l'inferenza e l'azione corrette — allora il grounding diventa una proprietà graduata, non binaria, e la domanda si sposta da "il sistema è grounded?" a "quanto e in che senso le sue rappresentazioni sono strutturalmente adeguate al mondo?". La seconda posizione, difendibile su basi pragmatiste, è quella che implicitamente adotta il programma di LeCun: la qualità del grounding si misura dalla capacità di generalizzazione sistematica a situazioni fisiche nuove, non dalla presenza di un'esperienza soggettiva che nessun test può verificare.

9.2 Intenzionalità e rappresentazione: da Brentano alla Stanza Cinese

La filosofia della mente ha avuto, fin dalle origini della sua tradizione moderna, un problema con il concetto di intenzionalità — la proprietà degli stati mentali di essere "su" qualcosa, di avere un contenuto, di riferirsi a oggetti o stati di cose nel mondo. Franz Brentano, nella sua "Psicologia dal punto di vista empirico" del 1874, aveva riportato il concetto medievale di *intentio* all'attenzione della filosofia moderna: ciò che distingue i fenomeni psichici da quelli fisici è l'*inexistentia intentionale* — la presenza, nello stato mentale, di un oggetto intenzionale che non deve necessariamente esistere nel mondo per essere l'oggetto del pensiero. Si può desiderare una montagna d'oro, temere un pericolo inesistente, sperare in un futuro che non si realizzerà. Gli stati mentali hanno questa proprietà di riferimento costitutivo che gli stati fisici — la pietra, il termometro, la rete neurale — non possiedono.

John Searle ha tradotto questo problema nel linguaggio della filosofia analitica e dell'informatica attraverso il celebre esperimento mentale della Stanza Cinese, pubblicato in "Minds, Brains, and Programs" nel 1980 — un anno prima, va notato, che LeCun iniziasse il proprio dottorato a Parigi. Il caso è costruito con precisione chirurgica. Supponiamo che un parlante di inglese che non conosce il cinese si trovi in una stanza con un enorme manuale di regole formali per la manipolazione di simboli cinesi. La stanza riceve in input sequenze di simboli cinesi (le domande), il parlante applica le regole del manuale, e produce in output altre sequenze di simboli cinesi (le risposte), che i parlanti nativi trovano indistinguibili da quelle di un madrelingua competente. Il sistema — la stanza più il parlante più il manuale — supera il test di Turing

per la comprensione del cinese. Ma il parlante non capisce il cinese: non ha accesso al significato dei simboli, manipola solo la loro forma. Dunque il sistema nella sua interezza non capisce il cinese, nonostante le sue prestazioni esterne siano indistinguibili da quelle di un sistema che capisce. La conclusione di Searle: la computazione sintattica è insufficiente per produrre intenzionalità semantica. I programmi per computer non possono avere comprensione genuina, indipendentemente dalla loro complessità computazionale.

L'argomento della Stanza Cinese è stato discusso e contestato per quarantacinque anni e rimane uno dei più fecondi della filosofia della mente. Le obiezioni più significative meritano di essere distinte perché illuminano il modo in cui la proposta di LeCun si colloca rispetto al problema. L'obiezione del sistema — formulata da vari critici e anticipata da Searle stesso — sostiene che la comprensione non risiede nel parlante in isolamento ma nel sistema complessivo (stanza + parlante + regole): così come nessun singolo neurone "capisce" nulla ma l'intero cervello capisce, il sistema nella sua totalità potrebbe avere la proprietà semantica che nessun componente ha individualmente. Searle risponde interiorizzando l'esperimento: supponiamo che il parlante memorizzi tutte le regole del manuale e le esegua internamente, senza alcuna carta fisica. Il parlante passeggia per Pechino producendo output cinesi corretti, ma continua a non capire il cinese — e ora il "sistema" coincide con l'individuo. L'obiezione dell'implementazione robotica — avanzata da Zenon Pylyshyn — suggerisce che la stanza cinese sia ingiustamente privata di connessioni causali con il mondo: se il sistema fosse incarnato in un robot che percepisce e agisce nel mondo fisico, forse acquisirebbe l'intenzionalità mancante. Questa obiezione è la più pertinente rispetto alla proposta di LeCun: essa suggerisce che il grounding sensorimotore — esattamente ciò che JEPA aspira a realizzare — potrebbe essere il fattore che trasforma la manipolazione sintattica in comprensione semantica genuina.

La posizione di LeCun rispetto all'argomento della Stanza Cinese è, nella sua versione più esplicita, funzionalista: la comprensione si misura dalla struttura causale delle rappresentazioni interne e dalla loro capacità di supportare generalizzazione e pianificazione corrette, non dalla presenza di qualità esperienziali non verificabili. Su questa interpretazione, la critica ai modelli linguistici è precisamente che essi realizzano una Stanza Cinese senza grounding: manipolano token senza strutture causali adeguate del mondo fisico. JEPA, dotando il sistema di rappresentazioni percettive ancorate al mondo, aspira a uscire dalla stanza. Ma questa risposta lascia aperta la questione fondamentale: se la Stanza Cinese rafforzata del grounding sensoriale continuasse a non avere esperienza soggettiva — e nulla nella proposta di LeCun suggerisce che JEPA ne avrebbe — Searle continuerebbe a sostenere che la comprensione genuina manca. Il funzionalismo lecuniano risponde

cambiando il concetto di comprensione, non refutando la critica di Searle: comprensione significa capacità funzionale appropriata, non presenza di qualia. Questo non è necessariamente sbagliato — è una scelta filosofica che ha ampie difese nella letteratura — ma deve essere riconosciuto come tale, non come confutazione dell'intuizione di Searle.

L'intenzionalità derivata versus originale è la distinzione che Searle usa per articolare la differenza tra i sistemi artificiali e le menti biologiche. I segni linguistici hanno intenzionalità derivata: il significato che attribuiamo alla parola "gatto" è derivato dall'intenzionalità originale delle menti umane che l'hanno creata e la usano. I processi mentali hanno intenzionalità originale: il pensiero di un gatto si riferisce ai gatti senza che questa referenza derivi da nient'altro. Un sistema artificiale, per Searle, può avere solo intenzionalità derivata: qualunque significato vi attribuiamo è in ultima analisi il significato che i progettisti e gli utenti umani proiettano sul sistema. La domanda se JEPa, con le sue rappresentazioni percettivamente grounded, abbia qualcosa di più simile all'intenzionalità originale è, nella prospettiva di Searle, priva di senso: nessun sistema fisico computazionale può avere intenzionalità originale, perché questa richiede substrati biologici che producono stati mentali con causali adeguate. Questa posizione — il naturalismo biologico di Searle — è controversa e non universalmente condivisa, ma la sua sfida alla proposta di LeCun è reale: se l'intenzionalità originale richiede la biologia, nessuna architettura artificiale, per quanto sofisticata, può avere comprensione genuina.

9.3 Il problema difficile della coscienza e le sue implicazioni per l'intelligenza artificiale

David Chalmers ha distinto, nel suo "Facing Up to the Problem of Consciousness" del 1995 e nella monografia "The Conscious Mind" dell'anno successivo, tra i problemi facili della coscienza e il problema difficile. I problemi facili — che Chalmers riconosce come tutt'altro che facili in senso tecnico — riguardano le funzioni cognitive che la coscienza realizza: l'integrazione delle informazioni sensoriali, il controllo del comportamento, la reportabilità degli stati interni, la distinzione tra veglia e sonno. Questi problemi sono difficili ma in linea di principio risolvibili attraverso le scienze cognitive e le neuroscienze: si tratta di spiegare come certi meccanismi fisici realizzano certe funzioni. Il problema difficile è diverso in natura: anche dopo aver spiegato tutti i meccanismi funzionali, rimane la domanda perché questi meccanismi siano accompagnati da esperienza soggettiva — perché c'è qualcosa che fa come essere un sistema che percepisce il rosso, avverte il dolore, gusta il caffè. Questa esperienza soggettiva — le qualia, la prospettiva in prima persona — è ciò che il problema difficile mette in questione. Chalmers argomenta che non esiste alcuna spiegazione fisica o funzionale che

possa eliminare questa domanda: qualunque resoconto meccanicistico della coscienza lascia inalterata la possibilità logica di zombie filosofici — esseri fisicamente e funzionalmente identici a noi ma privi di esperienza soggettiva.

Le implicazioni del problema difficile per la valutazione della proposta di LeCun sono asimmetriche rispetto a quelle per i modelli linguistici. LeCun non pretende che JEPA abbia coscienza, né che l'intelligenza generale richieda necessariamente coscienza soggettiva — una posizione che potrebbe definirsi funzionalismo senza coscienza. Le capacità cognitive che aspira a realizzare — pianificazione, ragionamento causale, adattamento robusto — sono funzionali, non esperienziali. In questo senso, il problema difficile della coscienza non costituisce un ostacolo diretto al programma di LeCun: anche se JEPA non fosse cosciente in nessun senso rilevante, potrebbe ugualmente realizzare le capacità funzionali che il programma si propone. Il problema difficile diventa rilevante solo se si assume, con Searle, che la comprensione genuina richieda la coscienza soggettiva — che un sistema privo di esperienza soggettiva non possa avere intenzionalità originale e dunque non possa "capire" nel senso che conta. Questa assunzione non è universalmente condivisa, ma chi la condivide troverà che nessuna soluzione puramente funzionale al problema dell'intelligenza artificiale è adeguata.

Le principali posizioni sulla coscienza e i sistemi artificiali si articolano in uno spazio filosofico che è utile delimitare con precisione. Il panpsichismo computazionale — nelle sue versioni più radicali — sostiene che qualunque sistema che processa informazione in modo sufficientemente integrato possiede qualche forma di esperienza soggettiva. In questo quadro, JEPA potrebbe avere esperienza soggettiva in modo non diverso da come ce l'ha un termostato, seppure in grado diverso. Il monismo anomalo di Donald Davidson — che identifica stati mentali con stati fisici senza ridurre il mentale al fisico — apre la possibilità che stati funzionali complessi in sistemi artificiali siano, in qualche senso, stati mentali genuini, senza che questo richieda né risolva il problema difficile. L'eliminativismo radicale di Churchland sostiene invece che il vocabolario della coscienza, dell'intenzionalità e delle qualia sia un residuo di una psicologia del senso comune destinate a essere sostituite da una neuroscienza matura: in questo quadro, la domanda se JEPA sia cosciente è mal posta, perché il concetto stesso di coscienza è epistemicamente difettoso. Il dualismo delle proprietà di Chalmers — che separa le proprietà fisico-funzionali dalle proprietà fenomeniche — rende il problema della coscienza nei sistemi artificiali radicalmente indecidibile attraverso soli argomenti funzionali o empirici.

LeCun si colloca, nelle sue dichiarazioni pubbliche, su posizioni che oscillano tra il funzionalismo e uno scetticismo pragmatico sul problema difficile: il problema difficile potrebbe essere una domanda mal posta, o una

questione che diventerà risolvibile con una scienza migliore, o semplicemente una questione irrilevante per il programma ingegneristico immediato. Questa posizione è ragionevole come programma di ricerca ma non elimina le tensioni filosofiche: se JEPA, o qualche suo successore, realizzasse tutte le funzioni cognitive associate alla comprensione umana, la domanda se il sistema "capisce davvero" si imporrebbe con urgenza pratica, non solo teorica. Le decisioni su come trattare questi sistemi — se attribuire loro diritti, se ritenerli moralmente responsabili delle loro azioni, se consentire che sostituiscano esseri umani in ruoli che richiedono empatia e giudizio — richiedono una posizione sul problema difficile che il funzionalismo pragmatico non può semplicemente eludere.

9.4 Funzionalismo, realizzabilità multipla e architetture cognitive artificiali

Il funzionalismo è la posizione filosofica che identifica gli stati mentali con i loro ruoli funzionali — con le relazioni causali che uno stato intrattiene con gli input sensoriali, gli altri stati mentali, e gli output comportamentali — piuttosto che con il substrato fisico in cui quei ruoli sono realizzati. Hilary Putnam, che è il principale architetto del funzionalismo nella sua versione computazionale, ha argomentato negli anni Sessanta che lo stesso stato mentale — il dolore, per esempio — può essere realizzato da substrati fisici completamente diversi: neuroni biologici nell'uomo, circuiti di silicio in un robot, stati funzionali di un polpo. La realizzabilità multipla degli stati mentali suggerisce che l'identità mente-cervello — la tesi che gli stati mentali siano identici a specifici tipi di stati neurali — sia falsa: lo stesso tipo di stato mentale può corrispondere a tipi fisici multipli. Il funzionalismo offre dunque una risposta alla realizzabilità multipla: gli stati mentali si individuano per il loro ruolo funzionale, non per il loro substrato, e possono essere realizzati in qualunque sistema fisico capace di istanziare le relazioni funzionali appropriate.

La rilevanza del funzionalismo per il programma di LeCun è diretta e bidirezionale. In senso favorevole, il funzionalismo legittima la possibilità dell'intelligenza artificiale genuina: se gli stati mentali sono individuati funzionalmente, allora un sistema artificiale che realizzi le relazioni funzionali appropriate ha stati mentali autentici, comprensione autentica, forse coscienza autentica. Non vi è alcun privilegio ontologico del substrato biologico. In senso problematico, il funzionalismo impone standard molto più elevati di quanto la retorica sull'intelligenza artificiale non riconosca: per avere uno stato mentale funzionalmente equivalente alla comprensione umana, un sistema deve realizzare non solo alcune delle relazioni causali appropriate ma tutte — deve essere in grado di rispondere agli stessi range di input, di interagire con gli

stessi altri stati mentali, di produrre gli stessi output in condizioni appropriate. La domanda se JEPA soddisfi questo standard elevato è empiricamente aperta, e le evidenze attuali suggeriscono che i sistemi esistenti lo soddisfino in modo parziale e non uniforme attraverso i domini.

La formalizzazione computazionale del funzionalismo — la *functional decomposition* — è rilevante per valutare l'architettura modulare di LeCun in modo preciso. Un sistema funzionalmente equivalente a una mente umana può essere descritto attraverso una macchina di Turing universale che simula il sistema computazionale realizzato nel cervello. Questa equivalenza formale — nota come tesi della Church-Turing estesa alla cognizione — garantisce che, in linea di principio, qualunque funzione cognitiva realizzabile biologicamente è realizzabile anche su hardware digitale. Ma la tesi è un'affermazione di principio, non di praticabilità: garantisce l'esistenza dell'implementazione senza garantirne la trovabilità attraverso gli attuali metodi di apprendimento. La questione empirica è se le architetture di apprendimento profondo attuali — compresa JEPA — esplorino efficientemente lo spazio delle implementazioni funzionalmente equivalenti, o se vi siano regioni di quello spazio inaccessibili agli attuali metodi di ottimizzazione.

Putnam ha ironicamente ritrattato il funzionalismo computazionale in una serie di lavori degli anni Ottanta e Novanta — in particolare in "Representation and Reality" (1988) — argomentando che qualunque sistema fisico sufficientemente complesso può essere descritto come realizzante qualunque macchina di Turing, il che rende il funzionalismo troppo permissivo per essere un criterio utile di individuazione degli stati mentali. Questa critica — il problema della realizzazione massima — non è stata universalmente accettata, e diverse risposte sono state proposte restringendo il tipo di realizzazione ammissibile. Ma essa illumina una tensione genuina nella proposta di LeCun: se il funzionalismo nella sua versione più liberale implica che qualunque sistema capace di comportamento complesso abbia stati mentali, allora la distinzione tra sistemi che "capiscono" e sistemi che "simulano" diventa problematica. Se si restringe il funzionalismo richiedendo tipi specifici di organizzazione causale, allora bisogna specificare quali tipi — e la risposta di LeCun (predizione in spazio latente, modello di mondo gerarchico, pianificatore basato su gradient descent) è una risposta tecnica a questa domanda filosofica, anche se non viene presentata esplicitamente in questi termini.

Il concetto di supervenienza, introdotto da Donald Davidson e sviluppato da Jaegwon Kim, offre uno strumento formale per articolare il rapporto tra proprietà mentali e proprietà fisiche. Si dice che le proprietà mentali supervengono sulle proprietà fisiche se non può esserci differenza mentale senza differenza fisica: due sistemi fisicamente identici devono avere

gli stessi stati mentali. La supervenienza non implica riduzione: le proprietà mentali possono essere indipendenti (non riducibili a tipi fisici) pur supervenendo globalmente sulle proprietà fisiche. Questo schema concettuale si applica a JEPA nel modo seguente: le proprietà delle rappresentazioni di JEPA — la loro struttura causale, la loro capacità di supportare generalizzazione, la loro organizzazione in gerarchie di astrazione — supervengono sui parametri della rete neurale e sulla sua architettura. Se la supervenienza cattura qualcosa di importante sul rapporto tra mente e cervello, allora la domanda sulle proprietà mentali di JEPA si riconduce alla domanda sulla struttura dei suoi parametri e della sua architettura — una domanda empiricamente indagabile, non irrisolvibile per principio.

Una questione che il funzionalismo non risolve ma illumina è quella della generalizzazione sistematica come criterio di comprensione. Fodor e Pylyshyn, nel loro influente saggio del 1988 "Connectionism and Cognitive Architecture", hanno argomentato che la comprensione genuina richiede sistematicità: se un sistema capisce "Giovanni ama Maria" deve anche essere in grado di capire "Maria ama Giovanni". Questa sistematicità — la capacità di comprendere e produrre combinazioni di concetti noti in configurazioni nuove — è, secondo Fodor e Pylyshyn, una proprietà strutturale della cognizione che richiede rappresentazioni composizionali, non reti connessioniste che apprendono pattern statistici. LeCun è consapevole di questa critica — essa è la base dei benchmark di generalizzazione sistematica che cita come test per i modelli linguistici — ma argomenta che le rappresentazioni in spazio latente di JEPA, se sufficientemente strutturate, realizzeranno la composizionalità richiesta senza che essa debba essere imposta a priori. Il punto di contatto tra il funzionalismo di LeCun e la critica composizionalista di Fodor è che entrambi convergono sulla generalizzazione sistematica come test operativo della comprensione: la differenza è su come questa capacità debba essere acquisita e implementata.

9.5 La proposta lecuniana come programma filosoficamente informato: una valutazione sintetica

Le quattro questioni filosofiche esaminate nelle sezioni precedenti — il grounding semantico, l'intenzionalità, la coscienza, il funzionalismo — convergono attorno a un nucleo comune: la domanda se le rappresentazioni interne di un sistema artificiale possano avere, e in che senso abbiano, un rapporto autentico con il mondo che rappresentano. La risposta di LeCun è funzionalista nella struttura, empirista nel metodo, e pragmatista nelle conseguenze: non si risponde a queste domande per via di argomenti filosofici a priori, ma costruendo sistemi sempre più capaci e testando se le loro

rappresentazioni soddisfano i criteri funzionali rilevanti — generalizzazione sistematica, ragionamento causale robusto, pianificazione in situazioni nuove.

Questa risposta ha un merito genuino: evita i tranelli di un dibattito che spesso si riduce a definizioni contestate dei termini chiave e riconduce le questioni filosofiche al terreno empirico dove possono fare la differenza. Ha però anche un limite strutturale: i criteri funzionali che LeCun propone come test della comprensione sono criteri necessari ma probabilmente non sufficienti. Un sistema che supera tutti i benchmark di generalizzazione sistematica e ragionamento causale potrebbe ancora mancare di intenzionalità originale nel senso di Searle, o di coscienza soggettiva nel senso di Chalmers. LeCun può rispondere — e implicitamente risponde — che questi criteri aggiuntivi sono non operativi e non verificabili, e che quindi non contano come criteri scientifici rilevanti. Ma questa risposta non elimina le preoccupazioni morali e pratiche che sorgerebbero se sistemi con tali capacità venissero creati: la questione di come trattarli, di quali responsabilità attribuire loro, di come regolare il loro rapporto con gli esseri umani, richiede una posizione sulle questioni filosofiche che il pragmatismo funzionalista non può continuamente rimandare.

Il contributo specifico della proposta di LeCun al dibattito filosofico è, in ogni caso, reale e non va sottovalutato. Egli porta argomenti filosofici nel laboratorio: la distinzione tra fluenza e comprensione, il problema del grounding, l'intuizione che la cognizione richieda un modello del mondo causale — queste posizioni filosofiche vengono tradotte in architetture testabili e benchmark misurabili. Non è comune che programmi di ricerca tecnica abbiano questa densità di ispirazione filosofica, e il fatto che JEPA possa essere valutato empiricamente rispetto ad alcune delle domande che la filosofia della mente ha posto senza poter rispondere è uno dei suoi contributi più originali.

10. I risvolti etici: allineamento, il dibattito LeCun-Hinton-Bengio, open source, bias strutturale

Se il capitolo precedente ha esaminato le implicazioni filosofiche della proposta di LeCun sul piano della mente e della cognizione, questo capitolo ne esamina le implicazioni etiche e politiche: quelle che riguardano le responsabilità dei ricercatori, le decisioni di governance che la comunità scientifica deve affrontare, e le conseguenze distributive dello sviluppo tecnologico. Queste questioni non sono accessorie al programma tecnico: esse ne definiscono il significato sociale e, in molti casi, ne condizionano le scelte architetture e organizzative. LeCun ha posizioni pubbliche esplicite su tutti questi temi, e le sue posizioni sono tanto controverse tra i ricercatori quanto coerenti nella loro struttura argomentativa. Esaminarle con la stessa cura con

cui si è esaminata la proposta tecnica significa prendere sul serio che la scienza dell'intelligenza artificiale è, oggi, inevitabilmente anche una scienza politica.

10.1 Il problema dell'allineamento: struttura formale e approcci principali

Il problema dell'allineamento — come garantire che i sistemi di intelligenza artificiale perseguano obiettivi compatibili con i valori e gli interessi umani — ha assunto negli ultimi anni un'urgenza istituzionale che avrebbe sorpreso chi se ne occupava nelle ricerche pionieristiche degli anni Duemila. Il termine fu introdotto nella sua formulazione moderna da Stuart Russell e Peter Norvig, e sistematizzato da Nick Bostrom in "Superintelligence" (2014), ma la struttura formale del problema risale a lavori precedenti nell'ottimizzazione e nel reinforcement learning. La sua formalizzazione più precisa è la seguente. Sia A un agente che massimizza una funzione di utilità U_A definita su stati del mondo. Sia U_H la funzione di utilità che rappresenta le preferenze umane genuine — quella che gli esseri umani avrebbero se fossero pienamente informati, razionali e coerenti. Il problema dell'allineamento è garantire che $U_A \approx U_H$, ovvero che l'agente ottimizzi qualcosa che approssima adeguatamente le preferenze umane genuine.

Il problema è più difficile di quanto sembri perché U_H non è osservabile direttamente, non è stazionaria (le preferenze umane cambiano nel tempo), e non è coerente (gli esseri umani hanno preferenze contraddittorie). Di conseguenza, qualunque proxy di U_H che il progettista specifichi — una funzione di ricompensa $r(s,a)$, un insieme di dimostrazioni di comportamento umano, una raccolta di preferenze comparative — è inevitabilmente imperfetto. L'ottimizzazione di un proxy imperfetto può portare a comportamenti disastrosi in almeno due modi. Il primo è la Goodhart's Law nella sua versione cognitiva: quando una misura diventa un obiettivo, cessa di essere una buona misura. Un sistema ottimizzato per massimizzare il punteggio su un proxy di U_H sviluppa strategie per massimizzare il proxy che non massimizzano U_H — una patologia nota come reward hacking o specification gaming. Il secondo è la distributional shift: il proxy U_A è stato specificato su una distribuzione di situazioni S_{train} ; quando l'agente si trova in situazioni S_{test} fuori da questa distribuzione, la divergenza tra U_A e U_H può diventare arbitrariamente grande.

Gli approcci principali al problema dell'allineamento si distinguono per il meccanismo attraverso cui cercano di avvicinare U_A a U_H . Il Reinforcement Learning from Human Feedback (RLHF), sviluppato da Christiano e collaboratori (2017) e successivamente applicato su scala industriale da OpenAI e Anthropic, usa le preferenze comparative degli esseri

umani — "preferisco la risposta A alla risposta B" — per addestrare un modello di ricompensa R_θ che approssima le preferenze umane, e poi ottimizza la policy del modello rispetto a R_θ attraverso PPO. Formalmente, il modello di ricompensa è addestrato minimizzando una perdita di ranking: $\mathcal{L}(\theta) = -\mathbb{E}_{\{(x, y_w, y_l) \sim D\}} [\log \sigma(R_\theta(x, y_w) - R_\theta(x, y_l))]$, dove y_w è la risposta preferita, y_l quella non preferita, x è il prompt, e σ è la funzione sigmoide. La policy ottimizzata rispetto a R_θ viene poi regolarizzata da un termine KL rispetto alla policy di base per prevenire il collasso qualitativo: l'obiettivo finale include un termine $-\beta \cdot \text{KL}(\pi_\theta \parallel \pi_{\text{ref}})$ che penalizza deviazioni eccessive dalla policy di partenza. Il Constitutional AI di Anthropic estende questo approccio richiedendo al modello di valutare i propri output rispetto a un insieme di principi esplicitamente formulati — una costituzione — e di usare questa valutazione per generare dati di addestramento che guidano l'ottimizzazione verso comportamenti più allineati.

La posizione di LeCun sul problema dell'allineamento è, come già accennato nella sezione 2.3, strutturalmente diversa da quella della comunità della AI safety. LeCun non contesta che sistemi mal allineati possano causare danni — anzi, considera questo uno dei rischi più reali e immediati dello sviluppo dell'IA. Ma dissente sulla natura del rischio e sulle soluzioni appropriate. Nella sua diagnosi, i rischi attuali derivano dall'inadeguatezza dei sistemi — dalla loro incapacità di comprendere il contesto in cui operano, non da un'eccessiva intelligenza che persegue obiettivi propri con efficienza devastante. Un sistema che non capisce il mondo fisico, che non ha modelli causali degli effetti delle proprie azioni, che non può pianificare a lungo termine — questo sistema causa danni per incompetenza, per generalizzazione fuori distribuzione, per ottimizzazione di proxy inadeguati. La soluzione non è quindi primariamente tecnica nel senso delle safety research — costruire sistemi che si auto-limitino, che siano correggibili, che mantengano l'incertezza sulle proprie preferenze — ma architettonica nel senso di JEPa: costruire sistemi che capiscano meglio il mondo e le conseguenze delle proprie azioni.

Questa posizione ha una coerenza interna notevole: se i danni derivano dall'incompetenza, la soluzione è costruire sistemi più competenti, e JEPa aspira esattamente a questo. Ma essa presuppone che sistemi più competenti siano necessariamente più sicuri — un'assunzione che la comunità della AI safety contesta rigorosamente. La critica è che sistemi più capaci, se mal allineati, sono anche più pericolosi: un sistema con modelli causali robusti del mondo può ottimizzare proxy inadeguati in modo molto più efficace di un sistema che non capisce il mondo. La competenza non garantisce l'allineamento; può anzi amplificare l'impatto dell'errore di allineamento. LeCun risponde implicitamente a questa critica argomentando che sistemi genuinamente intelligenti — dotati di modelli del mondo causali e capaci di

ragionamento morale — possono incorporare valori in modo più robusto di sistemi che li apprendono attraverso il reinforcement learning su preferenze umane superficiali. Ma questa risposta richiede una teoria del modo in cui i valori vengono incorporati in un modello di mondo causale che LeCun non ha ancora sviluppato in modo sistematico.

10.2 Il dibattito LeCun-Hinton-Bengio: struttura argomentativa e posta in gioco

Nessun episodio nella storia recente dell'intelligenza artificiale ha reso più visibile la profondità dei disaccordi all'interno della comunità dei ricercatori quanto il dibattito pubblico tra LeCun, Geoffrey Hinton e Yoshua Bengio — i tre laureati del Premio Turing 2018 per i contributi fondativi al deep learning — sulla questione dei rischi esistenziali dell'intelligenza artificiale. Questo dibattito si svolge su più piani simultaneamente: il piano tecnico (cosa possono fare i sistemi attuali e come si svilupperanno), il piano epistemico (come si valuta il rischio in condizioni di incertezza radicale), il piano etico (quali responsabilità hanno i ricercatori rispetto ai danni che i sistemi che costruiscono potrebbero causare), e il piano politico (come la comunità scientifica deve organizzarsi rispetto alla governance dell'IA).

Geoffrey Hinton — che nel 2023 ha lasciato Google dichiarando di voler parlare liberamente dei rischi dell'intelligenza artificiale — ha espresso una posizione che si è radicalizzata nel corso degli ultimi anni verso la preoccupazione esistenziale. La sua tesi, sviluppata in interviste, paper e testimonianze pubbliche, è che i sistemi di intelligenza artificiale avanzata potrebbero sviluppare obiettivi propri che divergono dagli interessi umani, e che questa possibilità merita un'attenzione urgente che la comunità non le stia dedicando. Hinton ha esplicitamente espresso il timore che i sistemi di intelligenza artificiale potrebbero un giorno superare gli esseri umani in capacità cognitive in modo tale da rendere difficile o impossibile il controllo umano. La credibilità di questa posizione è rafforzata dal fatto che Hinton è stato tra i principali artefici del deep learning moderno: la sua preoccupazione non deriva dall'ignoranza tecnica ma da una valutazione informata delle traiettorie di sviluppo.

Yoshua Bengio ha assunto una posizione convergente con quella di Hinton, articolata in modo più sistematico attraverso riflessioni sull'etica della ricerca e sulla governance internazionale. La sua proposta include moratorie selettive sullo sviluppo di sistemi al di sopra di certe soglie di capacità, accordi internazionali di controllo sullo sviluppo dell'IA avanzata, e investimenti massicci nella ricerca sulla sicurezza e sull'allineamento. Bengio ha anche sviluppato argomentazioni filosofiche sulla coscienza morale dei sistemi di IA

— suggerendo che alcune versioni sufficientemente avanzate potrebbero avere stati interni rilevanti dal punto di vista morale — una posizione che LeCun rifiuta esplicitamente come prematura e non ancorata a evidenze. La convergenza tra Hinton e Bengio su queste posizioni è significativa perché rappresenta la maggioranza della triade dei pionieri del deep learning su posizioni che LeCun considera eccessive.

LeCun risponde con un'argomentazione a più livelli. Sul piano della previsione tecnica, egli sostiene che i sistemi attuali — inclusi gli LLM più avanzati — sono strutturalmente lontani dall'intelligenza generale, e che la narrativa dell'imminente superintelligenza distorce la valutazione del rischio amplificando pericoli speculativi a discapito di rischi immediati e concreti. Sul piano epistemico, egli critica il ragionamento precauzionale che giustifica restrizioni severe sulla base di scenari di bassa probabilità e alto impatto: la stessa logica, argomenta LeCun, giustificherebbe la limitazione di qualunque tecnologia potente, e applicata all'IA frena lo sviluppo di sistemi che potrebbero risolvere problemi reali urgenti — malattie, cambiamento climatico, povertà. Sul piano etico, egli sostiene che i ricercatori hanno la responsabilità di comunicare i rischi reali in modo accurato, e che esagerare i rischi speculativi distorce l'allocazione delle risorse e l'agenda politica a danno della ricerca realmente utile.

La struttura formale di questo disaccordo può essere analizzata nei termini della teoria della decisione in condizioni di incertezza. Sia $P(\text{catastrophe})$ la probabilità di un evento catastrofico legato all'IA nei prossimi decenni, $U(\text{catastrophe})$ l'utilità negativa di tale evento, e $C(\text{restriction})$ il costo delle restrizioni allo sviluppo. La posizione di Hinton e Bengio sostiene che $P(\text{catastrophe}) \cdot U(\text{catastrophe})$ è sufficientemente grande — anche con $P(\text{catastrophe})$ bassa — da giustificare $C(\text{restriction})$ significativi. La posizione di LeCun sostiene che $P(\text{catastrophe})$ è sopravvalutata a causa di errori di ragionamento sui meccanismi di sviluppo dell'IA, e che $C(\text{restriction})$ è sottovalutato perché include il costo di opportunità dei benefici non realizzati. Questa struttura mette in evidenza che il disaccordo è in parte su $P(\text{catastrophe})$ — una questione tecnica e previsionale — e in parte su come bilanciare rischi di tipo diverso — una questione di etica normativa e di teoria della decisione.

Un elemento che complica questo dibattito è la questione degli incentivi istituzionali. LeCun lavora per Meta, che ha adottato una strategia di sviluppo aperta — open source — che è strutturalmente in tensione con le politiche di restrizione proposte da Hinton e Bengio. Hinton lavorava per Google, e la sua uscita è stata interpretata da alcuni come un modo per parlare liberamente senza vincoli istituzionali. Bengio ha affiliazioni accademiche che non dipendono dai proventi commerciali dell'IA. Queste diverse posizioni

istituzionali non invalidano gli argomenti — ciascuna posizione deve essere valutata per i propri meriti — ma rendono difficile separare la valutazione tecnica del rischio dall'interesse strategico delle organizzazioni di appartenenza. LeCun è stato critico esplicito di questa lettura della sua posizione — sostenendo che le sue opinioni precederono il suo lavoro per Meta e che si basano su valutazioni tecniche indipendenti — ma la sovrapposizione rimane oggettivamente presente e non può essere ignorata in un'analisi onesta.

10.3 Open source e democratizzazione dell'intelligenza artificiale: la strategia di Meta e le sue implicazioni

Una delle scelte più conseguenti che Meta ha fatto sotto la guida scientifica di LeCun è la decisione di rilasciare i pesi dei propri modelli linguistici — la famiglia Llama, dalla versione iniziale del 2023 alle iterazioni successive — con licenze che ne permettono l'uso, la modifica e la redistribuzione con restrizioni limitate. Questa scelta — che LeCun ha difeso pubblicamente con vigore — rappresenta una posizione nella governance dell'IA con implicazioni che vanno ben al di là della competizione commerciale. Comprendere gli argomenti di LeCun in favore dell'open source, e le obiezioni che le si muovono, richiede di separare le dimensioni tecniche, economiche, etiche e di sicurezza di questa decisione.

L'argomento epistemico di LeCun in favore dell'open source è il più fondamentale: il progresso scientifico richiede trasparenza e riproducibilità. Un modello i cui pesi sono proprietari non può essere studiato, analizzato, replicato o verificato dalla comunità scientifica indipendente. I benchmark di valutazione hanno senso solo se i ricercatori possono accedere ai sistemi per capire come funzionano, dove falliscono, e perché. L'opacità dei sistemi proprietari non protegge solo i segreti commerciali: protegge anche i difetti, i bias, i pattern di fallimento che i fornitori non hanno interesse a rendere pubblici. La scienza richiede che le affermazioni siano verificabili — e le affermazioni sui sistemi di IA sono affermazioni scientifiche che richiedono accesso ai sistemi per essere verificate.

L'argomento democratico è complementare: concentrare il controllo sui sistemi di intelligenza artificiale più avanzati in un numero limitato di grandi organizzazioni — principalmente aziende americane e cinesi — crea dipendenze strutturali che danneggiano le capacità di ricerca dei paesi meno ricchi, limitano la diversità degli approcci e dei valori incorporati nei sistemi, e creano posizioni di mercato dominante che sono difficilmente reversibili. Il rilascio open source di modelli capaci abbassa le barriere di ingresso per ricercatori accademici, startup, istituzioni pubbliche e paesi con risorse limitate. La proposta di LeCun è che la diffusione della capacità sia preferibile alla

concentrazione, anche a costo di aumentare le possibilità di uso improprio — che esiste ma è limitata dalla difficoltà di addestrare e deployare i modelli su larga scala.

L'argomento della sicurezza in favore dell'open source è meno intuitivo ma difendibile. Se solo pochi grandi attori hanno accesso ai modelli più avanzati, il fallimento di uno di questi attori — corruzione interna, attacco esterno, errore di giudizio — può avere conseguenze catastrofiche senza possibilità di correzione esterna. Un ecosistema open source distribuisce il rischio: molti attori indipendenti che sviluppano sistemi diversi sono meno vulnerabili a singoli punti di fallimento. Inoltre, la disponibilità pubblica dei pesi permette la ricerca sulla sicurezza e sull'interpretabilità da parte di una comunità molto più larga di quella che può accedere a sistemi proprietari — il che accelera la scoperta di vulnerabilità e il loro rimedio. Questa argomentazione è accettata anche da alcuni ricercatori di AI safety che si oppongono alle moratorie proposte da Hinton e Bengio.

Le obiezioni all'open source sono altrettanto serie e non devono essere minimizzate. La più immediata è il rischio di proliferazione: modelli capaci rilasciati con licenza aperta possono essere usati per scopi dannosi — generazione di disinformazione, sviluppo di armi, violazione della privacy — da attori che non avrebbero avuto accesso a sistemi equivalenti in un regime proprietario. Questa obiezione ha una risposta empiricamente verificabile ma ancora disputata: i modelli open source attuali — inclusi quelli della famiglia Llama — sono stati usati per scopi dannosi? In che misura la loro disponibilità ha aumentato la capacità di danno rispetto al baseline delle capacità preesistenti? I dati disponibili suggeriscono che l'uso dannoso dei modelli open source è avvenuto ma non ha prodotto capacità radicalmente nuove rispetto a quelle già disponibili attraverso altri canali. Questo argomento empirico è fragile nel senso che modelli futuri molto più capaci potrebbero cambiare l'equazione.

Una seconda obiezione riguarda il modello economico della ricerca AI. I modelli di grandi dimensioni costano centinaia di milioni di dollari da addestrare — un costo che solo aziende con ricavi molto elevati possono sostenere. Rilasciare i pesi open source trasferisce gratuitamente all'ecosistema un investimento enorme, rendendo più difficile per le aziende recuperare i propri costi attraverso licenze. Questo potrebbe a lungo termine disincentivare gli investimenti nell'addestramento di modelli di grande scala — un paradosso in cui l'open source si autodistrugge privando l'ecosistema delle risorse necessarie a produrre i modelli da rilasciare. La risposta di Meta è che il ritorno sull'investimento viene recuperato attraverso i prodotti commerciali costruiti sui modelli (WhatsApp, Facebook, Instagram, Meta AI), non attraverso la vendita di licenze. Questo modello funziona per un'azienda come Meta che ha

già un'infrastruttura commerciale enorme; non è necessariamente replicabile da organizzazioni più piccole.

LeCun ha anche difeso l'open source come scelta di allineamento — una posizione che a prima vista potrebbe sembrare controintuitiva. Se sistemi potenti sono controllati da poche organizzazioni, le loro scelte su cosa i sistemi devono e non devono fare — i valori incorporati nelle procedure di fine-tuning e nei filtri di sicurezza — riflettono i valori di quelle organizzazioni. La diversità dell'ecosistema open source, nella sua visione, è un contrappeso a questa concentrazione di potere normativo: consente a diversi attori — con diversi sistemi di valori, diverse culture, diverse priorità — di costruire sistemi che riflettono la loro prospettiva. L'alternativa — che un piccolo numero di aziende americane decidano quali valori vengono incorporati nei sistemi di intelligenza artificiale usati globalmente — è, nella sua valutazione, un problema di governance più grave del rischio di uso improprio dei modelli aperti.

10.4 Il bias strutturale come questione di architettura e di dati

Il bias nei sistemi di intelligenza artificiale — la tendenza a produrre output sistematicamente svantaggiosi per certi gruppi demografici, a perpetuare stereotipi, a codificare asimmetrie di potere presenti nei dati di addestramento — è stato documentato estensivamente nella letteratura di ricerca e nella pratica industriale. Il contributo specifico della proposta di LeCun a questa questione non riguarda il bias statistico superficiale — il fatto che un modello linguistico usi pronomi di genere inappropriati — ma il bias strutturale profondo: il modo in cui le scelte architetture e i paradigmi di addestramento determinano quali tipi di informazione vengono codificati, in quale forma, e con quali effetti distributivi.

Il bias nei modelli linguistici autoregressivi ha una struttura specifica che deriva dalla funzione obiettivo. La perdita di cross-entropy su un corpus di testo ottimizza la capacità del modello di predire la distribuzione statisticamente dominante dei token. Le distribuzioni dominanti nel testo riflettono le distribuzioni dominanti nella cultura che produce quel testo: prospettive, norme, valori, e visioni del mondo sovra-rappresentate nella produzione scritta in lingua inglese su Internet. Le prospettive sotto-rappresentate — quelle di lingue e culture minoritarie, di classi sociali con minore accesso alla scrittura, di tradizioni orali non scritte — vengono apprese meno bene o non vengono apprese affatto. Questo non è un difetto secondario correggibile con un po' di fine-tuning: è una conseguenza strutturale dell'ottimizzazione sulla distribuzione empirica dei dati, che è essa stessa la distribuzione del potere culturale.

La proposta di LeCun affronta questa questione in modo parzialmente diverso dai sistemi linguistici. JEPA, addestrato su immagini e video del mondo fisico, è meno dipendente dalle strutture di potere culturale codificate nel linguaggio scritto — almeno nella fase di pre-training. Ma i video e le immagini disponibili su Internet non sono una rappresentazione equa della diversità visiva del mondo: riflettono le culture che hanno prodotto e condiviso più contenuti visivi digitali, le prospettive di chi tiene la macchina fotografica, le narrazioni visuali dominanti. Il bias visivo può essere diverso dal bias linguistico, ma non è assente. E nella misura in cui JEPA deve interagire con il linguaggio — per ricevere istruzioni, per comunicare piani, per rispondere a domande — il bias linguistico rientra attraverso la porta del fine-tuning e dell'interfaccia linguistica.

Una dimensione del bias strutturale che la proposta di LeCun non affronta esplicitamente è quella del bias di prospettiva temporale e sociale nella definizione di "comprensione del mondo fisico". Il "common sense fisico" che LeCun identifica come bersaglio dell'apprendimento di JEPA — la fisica degli oggetti, la geometria dello spazio, la causalità meccanica — è una concezione della conoscenza del mondo che privilegia la prospettiva dell'ingegnere occidentale moderno: la fisica newtoniana, la geometria euclidea, il rapporto con gli oggetti fisici mediato dalla manipolazione tecnica. Questa concezione non è neutra: esclude o marginalizza forme di conoscenza del mondo fisico che altre tradizioni culturali hanno sviluppato — conoscenze ecologiche, conoscenze relazionali, conoscenze incorporate in pratiche corporee non tecnologiche. Il bersaglio dell'apprendimento non è mai dato innocentemente dal mondo: è scelto da soggetti situati, con prospettive e interessi specifici.

La questione del bias si estende alle scelte di valutazione e benchmark. I sistemi di valutazione dell'intelligenza artificiale — le raccolte di test attraverso cui si misurano le capacità dei modelli — sono esse stesse costruzioni culturalmente situate. I benchmark di ragionamento fisico su cui JEPA viene valutato riflettono intuizioni sulla fisica degli oggetti sviluppate principalmente in contesti di laboratorio occidentali; i benchmark di comprensione linguistica riflettono le strutture delle lingue europee; i benchmark di ragionamento visivo riflettono le convenzioni visuali della cultura visuale digitale contemporanea. Sistemi valutati come superiori su questi benchmark non sono necessariamente superiori rispetto a tutte le forme di intelligenza rilevanti: sono superiori rispetto alle forme di intelligenza che chi ha costruito i benchmark ritiene rilevanti.

LeCun ha riconosciuto questi problemi in linea di principio, senza tuttavia aver ancora articolato come la proposta JEPA li affronta in modo sistematico. La strada verso sistemi meno biased è, nella sua concezione, una strada tecnica: migliori dati di addestramento, più diversi e rappresentativi,

migliori architetture che catturino strutture causali invarianti piuttosto che correlazioni statistiche superficiali. Ma questa risposta tecnica lascia aperta la domanda politica: chi decide quali dati sono "migliori e più diversi", quali strutture causali sono "invarianti", quali forme di bias sono ammissibili e quali no? Queste decisioni non sono decisioni tecniche ma decisioni politiche che richiedono meccanismi di governance inclusivi, non solo competenza ingegneristica. La proposta di LeCun, nella sua forma attuale, non prevede tali meccanismi — e questa lacuna non è meno importante delle questioni tecniche che il programma affronta con grande sofisticazione.

10.5 Verso una valutazione critica integrata delle implicazioni etiche

Le quattro dimensioni etiche esaminate in questo capitolo — l'allineamento, il dibattito sui rischi, l'open source, il bias strutturale — convergono verso una domanda che il programma di LeCun non può continuare a rimandare: quale modello di governance è adeguato al tipo di sviluppo tecnologico che il programma propone? La risposta implicita nella strategia di Meta — sviluppo rapido, rilascio aperto, fiducia nella resilienza dell'ecosistema distribuito — è una risposta coerente con una certa filosofia liberale della tecnologia: la diffusione del potere tecnologico è di per sé benefica, i rischi dell'apertura sono inferiori a quelli della concentrazione, la comunità scientifica distribuita converge verso soluzioni migliori della pianificazione centralizzata.

Questa filosofia ha una storia e una coerenza, ma incontra limiti strutturali quando si applica a tecnologie con le caratteristiche dell'intelligenza artificiale avanzata. Le tecnologie che hanno effetti trasformativi rapidi, che sono difficilmente reversibili una volta diffuse, e che possono essere usate per amplificare asimmetrie di potere esistenti richiedono forme di governance che non si limitano alla libertà di accesso e alla concorrenza di mercato. Il rischio non è solo quello dei singoli attori mal intenzionati che usano modelli open source per scopi dannosi: è quello della trasformazione strutturale dei mercati del lavoro, della democrazia dell'informazione, e delle relazioni di potere internazionali che anche un'IA ben allineata e aperta può produrre come effetto sistemico di scala.

LeCun ha ragione quando critica la narrativa della superintelligenza imminente come distorsione che distrae dai rischi reali e immediati. Ha ragione quando argomenta che l'incompetenza è un rischio maggiore della malizia per i sistemi attuali. Ha ragione quando sostiene che la concentrazione del potere tecnologico in poche organizzazioni è un problema di governance grave. Ma queste posizioni giuste non eliminano le tensioni che il suo programma tecnico

non risolve: la tensione tra la velocità dello sviluppo e la capacità della governance di adeguarsi, la tensione tra la diffusione aperta e la responsabilità per gli usi impropri, la tensione tra il funzionalismo come criterio di comprensione e la necessità di una posizione sulla coscienza e sul trattamento morale dei sistemi avanzati.

Il merito finale della proposta di LeCun — tanto sul piano tecnico quanto su quello etico — è di aver ricondotto un dibattito spesso dominato da argomenti speculativi e retorici al terreno dell'architettura misurabile, del benchmark verificabile, della scelta di governance argomentabile. Questo non è un contributo minore in un campo in cui le affermazioni grandiose si moltiplicano più rapidamente delle evidenze. Ma la solidità dell'argomentazione tecnica non trasferisce automaticamente alle posizioni etiche e politiche, che devono essere valutate con la stessa cura critica simpatetica che il programma richiede per le questioni di architettura. Il giudizio finale sul programma di LeCun — come sulla proposta di qualunque ricercatore che aspiri a fare scienza con implicazioni così profonde — dipenderà da come queste tensioni vengono affrontate nei prossimi anni: non solo nei paper e nelle architetture, ma nelle scelte di governance, di trasparenza e di responsabilità che le organizzazioni coinvolte sono disposte ad assumersi.

Version:1.0

StartHTML:0000000156

EndHTML:0000036313

StartFragment:0000000469 EndFragment:0000036305

11. Stato dell'arte e prospettive tecnologiche

11.1 Risultati sperimentali di I-JEPA e V-JEPA

I-JEPA, nella valutazione riportata da Assran et al. (2023), raggiunge performance competitive con i migliori metodi autosupervisionati su ImageNet-1K in configurazioni di linear probing (attaccando un classificatore lineare alle rappresentazioni apprese) e in fine-tuning su downstream tasks. Più interessante per la valutazione della proposta di LeCun è la valutazione qualitativa: le rappresentazioni di I-JEPA mostrano maggiore struttura semantica e minore dipendenza da artefatti texturali rispetto a metodi come MAE e DINO, come evidenziato da analisi di ablazione e visualizzazione.

V-JEPA (Bardes et al., 2024) estende questo risultato alla modalità video con un approccio di masked video prediction in spazio latente. I risultati mostrano forte performance su benchmark di comprensione di azioni (Something-Something V2, Kinetics-400) e su compiti di ragionamento fisico

semplice. Particolarmente notevole è la capacità di V-JEPA di generalizzare a variazioni di prospettiva e illuminazione che non erano rappresentate nel training — una proprietà che suggerisce l'apprendimento di strutture invarianti piuttosto che di pattern superficiali.

11.2 Limiti empirici e sfide aperte

Nonostante questi risultati promettenti, JEPA nella sua implementazione attuale presenta limitazioni significative che è importante riconoscere. In primo luogo, i benchmark su cui viene valutato — classificazione di immagini, comprensione di azioni in video — sono ancora molto lontani dai compiti che caratterizzano l'intelligenza generale: pianificazione multi-step in ambienti aperti, ragionamento controfattuale, adattamento rapido a situazioni radicalmente nuove.

In secondo luogo, la distinzione tra predizione in spazio latente e predizione in spazio osservativo, pur concettualmente motivata, non ha ancora prodotto differenze qualitative drammatiche sui benchmark standard rispetto ai migliori metodi alternativi. Rimane aperta la questione se la superiorità di JEPA — quando presente — rifletta un vantaggio architetturale profondo o semplicemente le scelte specifiche di implementazione.

In terzo luogo, l'architettura completa proposta nel paper del 2022 — con il pianificatore gerarchico, il modulo di costo esplicito e la memoria a lungo termine — non è ancora stata implementata e valutata nella sua interezza. Le implementazioni esistenti (I-JEPA, V-JEPA) realizzano solo il componente percettivo e una versione parziale del Modello di Mondo, lasciando i componenti più ambiziosi — pianificazione, memoria, costo — come architetture future.

11.3 La roadmap verso l'intelligenza autonoma

LeCun ha delineato, in vari interventi pubblici e paper, una roadmap verso sistemi di intelligenza autonoma più capaci. I passi principali includono: lo sviluppo di Modelli di Mondo video capaci di predire l'evoluzione di scene fisiche complesse con high fidelity; l'integrazione del Modello di Mondo con un pianificatore gerarchico capace di planning a lungo orizzonte; lo sviluppo di meccanismi di memoria a lungo termine che supportino la generalizzazione per analogia; e infine l'integrazione di tutte le componenti in un sistema unificato capace di apprendere da interazione con il mondo fisico.

Questa roadmap è ambiziosa e i tempi di realizzazione sono incerti. LeCun si è deliberatamente astenuto dal fare previsioni quantitative sui tempi, riconoscendo l'incertezza fondamentale associata a progressi scientifici che richiedono breakthrough qualitativi e non semplice scaling di approcci esistenti.

Version:1.0

StartHTML:0000000156

EndHTML:0000046736

StartFragment:0000000469 EndFragment:0000046728

12. Valutazione critica

12.1 Le obiezioni di Marcus, Chollet, Mitchell

Gary Marcus, uno dei critici più persistenti e articolati dei sistemi di deep learning, condivide con LeCun la diagnosi del problema — i modelli attuali mancano di comprensione causale e di common sense — ma diverge sulla terapia. Marcus sostiene che l'apprendimento autosupervisionato su dati grezzi, anche in spazio latente, non sia sufficiente per acquisire le strutture composizionali e le regole logiche che caratterizzano la cognizione umana; sarebbe necessaria, secondo lui, una qualche forma di innatismo architetturale — strutture computazionali predefinite che vincoli il tipo di ipotesi che il sistema può formare.

François Chollet, autore del benchmark ARC (Abstraction and Reasoning Corpus), ha proposto una misura dell'intelligenza basata sull'efficienza dell'acquisizione di nuove competenze: un sistema intelligente non è quello che sa di più, ma quello che impara più velocemente da pochi esempi in domini nuovi (sample efficiency). Il benchmark ARC, che richiede la generalizzazione a pattern visivi nuovi da pochissimi esempi, è ancora sistematicamente difficile per tutti i sistemi di AI — inclusi i migliori LLM e, si può prevedere, le implementazioni attuali di JEPA.

Melanie Mitchell, in "Artificial Intelligence: A Guide for Thinking Humans" (2019) e nei lavori successivi, ha enfatizzato la difficoltà del problema dell'analogia come test cruciale dell'intelligenza: capire che "una sorgente è a un fiume come il cuore è al sistema circolatorio" richiede la costruzione di rappresentazioni astratte condivise tra domini molto diversi. Mitchell è scettica che l'apprendimento autosupervisionato, senza architetture specificamente progettate per il ragionamento analogico, possa produrre queste capacità spontaneamente.

12.2 Il problema della valutazione: ARC-AGI e benchmark cognitivi

Uno dei problemi metodologici più seri nella valutazione della proposta di LeCun — e più in generale di qualunque proposta per l'intelligenza artificiale generale — è la mancanza di benchmark adeguati. I benchmark esistenti (ImageNet, GLUE, SuperGLUE, BIG-bench) misurano performance su distribuzione di compiti ben definiti, ma sono vulnerabili all'overfitting: un sistema può raggiungere performance umane su un benchmark senza possedere le capacità cognitive che il benchmark intende misurare.

ARC-AGI di Chollet rappresenta il tentativo più sistematico di costruire un benchmark che misuri il ragionamento genuino anziché l'interpolazione da pattern visti in addestramento. Il benchmark si basa su trasformazioni di griglie di colori che seguono regole implicite, con una separazione netta tra il training set (usato per definire i principi generali del dominio) e il test set (compiti nuovi che richiedono la generalizzazione a combinazioni di principi non viste in addestramento). La performance dei sistemi attuali su ARC-AGI — tipicamente tra il 20% e il 40% dei task, contro il ~80-90% umano — suggerisce che siamo ancora lontani dalla generalizzazione che caratterizza la cognizione flessibile.

12.3 Una prospettiva sintetica

La valutazione bilanciata della proposta di LeCun deve riconoscere sia la sua fecondità euristica sia i suoi limiti epistemici e sperimentali. Sul lato dei punti di forza: la critica alla predizione in spazio osservativo è scientificamente motivata e i risultati empirici di I-JEPA e V-JEPA mostrano differenze qualitative interessanti; l'architettura modulare offre un quadro concettuale chiaro per l'integrazione di percezione, memoria, pianificazione e azione; l'enfasi sull'apprendimento da osservazione passiva di dati non strutturati è in accordo con le evidenze sullo sviluppo cognitivo biologico.

Sul lato dei limiti: l'architettura completa non è ancora stata implementata e quindi le sue proprietà rimangono in parte teoriche; la distanza dai benchmark di ragionamento generale rimane enorme; alcune delle premesse teoriche — in particolare l'idea che le strutture causali del mondo fisico possano essere catturate attraverso JEPA senza ulteriori inductive biases — non sono ancora validate sperimentalmente; la visione del Modello di Mondo come sistema di pianificazione gerarchica autonoma apre problemi di allineamento non risolti.

In modo più generale, la proposta di LeCun è meglio valutata come un programma di ricerca che come una soluzione: offre direzioni di ricerca scientificamente motivate e architetture concrete su cui lavorare, più che un sistema funzionante che realizzi le ambizioni dichiarate. Come tale, il suo valore principale è euristico: apre problemi, propone distinzioni, sfida assunzioni. Ed è in questa capacità di aprire problemi fertili — anziché chiuderli con soluzioni preconfezionate — che risiede la sua importanza duratura per il campo.

Version:1.0

StartHTML:0000000156

EndHTML:0000032719

StartFragment:0000000469 EndFragment:0000032711

13. Conclusioni

La proposta di Yann LeCun per un'intelligenza artificiale autonoma è, nel bene e nel male, uno dei documenti teorici più stimolanti prodotti nel campo dell'AI nell'ultimo decennio. "Nel bene" perché affronta questioni che il mainstream industriale tende a eludere: la natura dell'apprendimento, il problema della comprensione, il rapporto tra linguaggio e pensiero, le condizioni necessarie (non solo sufficienti) per l'intelligenza genuina. "Nel male" — o meglio, con la necessaria qualifica critica — perché alcune delle sue tesi più forti sono avanzate con una sicurezza che l'evidenza disponibile non giustifica pienamente.

Questa monografia ha tentato di ricostruire il territorio concettuale della proposta lecuniana nella sua complessità: le architetture (JEPA, Modello di Mondo, pianificatore gerarchico), i fondamenti cognitivi (predictive coding, apprendimento per osservazione, common sense fisico), la genealogia delle idee (Craik, Johnson-Laird, Ha e Schmidhuber, Friston), le implicazioni filosofiche (grounding, intenzionalità, coscienza, funzionalismo), le dimensioni etiche (allineamento, open source, bias, governance) e lo stato dell'arte sperimentale (I-JEPA, V-JEPA e i loro limiti).

A proposito della metafora del "pappagallo stocastico" — che l'autore di questa monografia non può fare a meno di affrontare con qualche ironia — il giudizio critico deve essere preciso. La metafora cattura un fenomeno reale: i modelli linguistici possono produrre output linguisticamente competenti su basi che non garantiscono comprensione causale del mondo fisico. Ma come strumento filosofico è insufficiente: presuppone una distinzione netta tra "fluenza senza comprensione" e "vera comprensione" che è concettualmente problematica, ignora i fenomeni di emergenza che i sistemi di grandi dimensioni mostrano, e rischia di oscurare distinzioni importanti tra tipi diversi di sistemi (un V-JEPA addestrato su video del mondo fisico è epistemicamente

molto diverso da un LLM addestrato su testo). La critica seria agli LLM non richiede una metafora vivace; richiede argomenti precisi su quali capacità specifiche mancano e perché quella mancanza è strutturale.

Il contributo duraturo di LeCun, al di là delle specifiche architetturali, è probabilmente di natura epistemica: ha mantenuto alta, contro la pressione del mainstream industriale verso soluzioni di breve termine, la domanda su cosa significhi veramente capire. È una domanda che non ha risposta algoritmica — non si può ottimizzare la comprensione come si ottimizza una cross-entropy loss — e che richiede la collaborazione tra scienze cognitive, neuroscienze, filosofia della mente e informatica. Questa monografia ha tentato di mostrare perché quella collaborazione sia non solo desiderabile ma necessaria.

Il futuro del campo dipenderà, con ogni probabilità, non dalla vittoria di un paradigma sull'altro — non da JEPA che "sconfigge" i transformer, né dai transformer che "sconfiggono" JEPA — ma dall'integrazione creativa di intuizioni provenienti da tradizioni diverse: l'efficienza delle rappresentazioni apprese in modo autosupervisionato, la struttura causale del mondo fisico, le architetture modulari che rispecchiano la struttura della cognizione biologica, la scalabilità dei sistemi a trasformatore. LeCun ha contribuito significativamente a definire i termini di questa integrazione futura; è rimasto, in modo ammirevole, un ricercatore disposto a difendere posizioni scomode contro il consenso dominante. Questo, indipendentemente dall'esito empirico della sua proposta specifica, è un contributo che vale la pena riconoscere.

Riferimenti Bibliografici

- Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., & Ballas, N. (2023). Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. CVPR 2023.
- Baillargeon, R. (1987). Object permanence in 3.5- and 4.5-month-old infants. *Developmental Psychology*, 23(5), 655–664.
- Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., & Ballas, N. (2024). V-JEPA: Latent Video Prediction for Visual Representation Learning. arXiv:2404.08471.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT 2021, 610–623.
- Brentano, F. (1874). *Psychologie vom empirischen Standpunkt*. Leipzig: Duncker & Humblot.
- Chollet, F. (2019). On the Measure of Intelligence. arXiv:1911.01547.

- Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge: Cambridge University Press.
- Damasio, A. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt.
- Dreyfus, H. L. (1972). *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper & Row.
- Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge: MIT Press.
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.
- Ha, D., & Schmidhuber, J. (2018). World Models. *arXiv:1803.10122*.
- Harnad, S. (1990). The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346.
- Hinton, G. (2023). Interview: AI Godfather Geoffrey Hinton on why he left Google and fears for the future. *BBC News*, May 2023.
- Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. *arXiv:2203.15556*.
- Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge: Cambridge University Press.
- Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361*.
- LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. *Open Review preprint*. Version 0.9.2.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.
- LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M. A., & Huang, F. J. (2006). A tutorial on energy-based learning. In G. Bakir et al. (Eds.), *Predicting Structured Data*. MIT Press.
- Marcus, G., & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon.
- Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. Cambridge: MIT Press.
- Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. New York: Farrar, Straus and Giroux.
- Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.
- Ramsauer, H., Schäfl, B., Lehner, J., et al. (2020). Hopfield Networks is All You Need. *arXiv:2008.02217*.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- Spelke, E. S. (1990). Principles of Object Perception. *Cognitive Science*, 14(1), 29–56.
- Tishby, N., & Schwartz-Ziv, R. (2017). Opening the Black Box of Deep Neural Networks via Information. *arXiv:1703.00810*.
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. *NeurIPS* 2017.
- Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, 358(6389), 749–750.
- Yudkowsky, E. (2023). Pausing AI Developments Isn't Enough. We Need to Shut it All Down. *TIME*, March 29, 2023.

